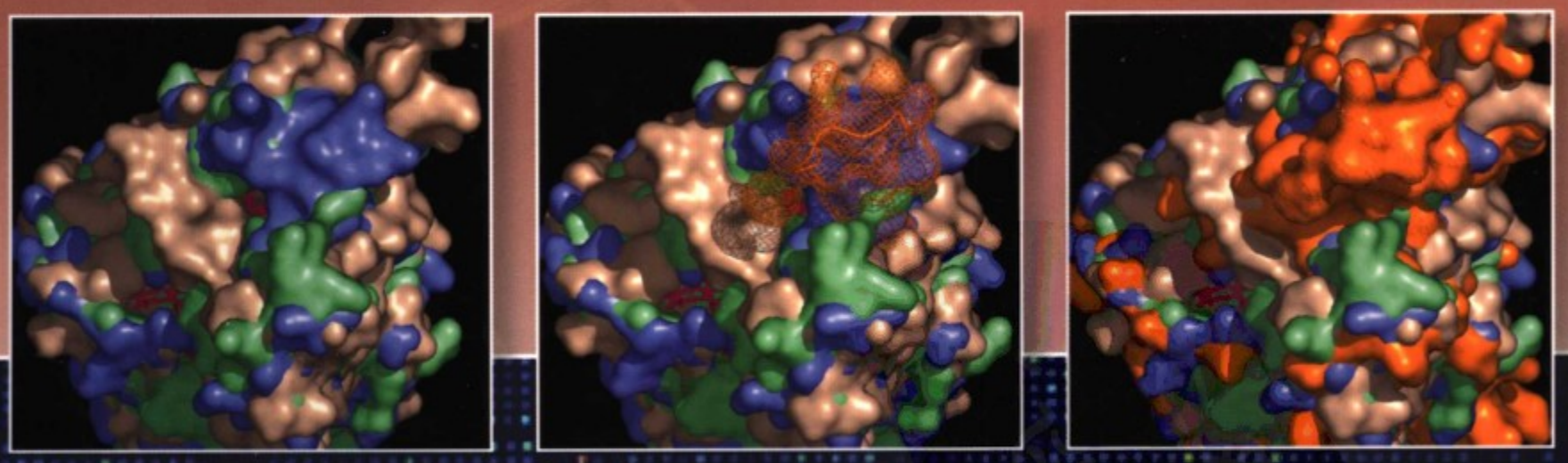
普通高等教育“十一五”国家级规划教材

# 现代分子生物学

# Modern Molecular Biology

（第3版）

朱玉贤　李　毅　郑晓峰　编著



高等教育出版社
Higher Education Press

# *Modern Molecular Biology*

*Third Edition*

ISBN 978-7-04-022214-2

9 787040 222142 >

定价 53.00 元

普通高等教育“十一五”国家级规划教材

# 现代分子生物学

# *Modern Molecular Biology*

（第3版）

朱玉贤　李　毅　郑晓峰　编著

高等教育出版社
Higher Education Press

## 内容提要

本书共分11部分，分别对染色体结构、DNA的复制形式与特点、DNA的转座、遗传密码的破译、蛋白质的合成和运转、基因表达调控的原理、癌症与癌基因活化、免疫缺损病毒(HIV)的分子机制等现代分子生物学中的重大问题做了全面系统的分析。其中第一、二章介绍了分子生物学、染色体与DNA的基本概念，第三至四章回顾了从DNA到RNA以及从mRNA到蛋白质的生物信息流，第七至八章叙述了参与原核、真核细胞基因表达调控的各种元件，探讨了DNA甲基化、蛋白质磷酸化、乙酰化修饰及各种不同环境因子对基因活性和功能的影响，第九至十章讨论了疾病与人类健康、基因与发育等重要生命现象的分子生物学基础，第十一章则讨论了基因组学与比较基因组学研究的最新成果。此外，本书还在第五、六两章讨论了主要分子生物学实验的技术和原理。

本书可供高等院校生物科学和生物技术专业的教师和学生使用，也可作为相关专业研究人员的参考书。

**图书在版编目(CIP)数据**

现代分子生物学/朱玉贤，李毅，郑晓峰编著. —3版.
北京：高等教育出版社，2007.11
ISBN 978-7-04-022214-2

Ⅰ.现… Ⅱ.①朱…②李…③郑… Ⅲ.分子生物学-高等学校-教材 Ⅳ.Q7

中国版本图书馆CIP数据核字(2007)第129636号

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 策划编辑 | 吴雪梅 | 责任编辑 | 薛 玥 | 封面设计 | 张 楠 | 责任绘图 | 尹 莉 |
| 版式设计 | 王 莹 | 责任校对 | 殷 然 | 责任印制 | 宋克学 | | |

| | | | |
|---|---|---|---|
| 出版发行 | 高等教育出版社 | 购书热线 | 010-58581118 |
| 社 址 | 北京市西城区德外大街4号 | 免费咨询 | 800-810-0598 |
| 邮政编码 | 100011 | 网 址 | http://www.hep.edu.cn |
| 总 机 | 010-58581000 | | http://www.hep.com.cn |
| | | 网上订购 | http://www.landraco.com |
| | | | http://www.landraco.com.cn |
| 经 销 | 蓝色畅想图书发行有限公司 | 畅想教育 | http://www.widedu.com |
| 印 刷 | 高等教育出版社印刷厂 | | |
| 开 本 | 850×1168 1/16 | 版 次 | 1997年3月第1版<br>2007年11月第3版 |
| 印 张 | 31 | 印 次 | 2007年11月第1次印刷 |
| 字 数 | 640 000 | 定 价 | 53.00元 |

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换。

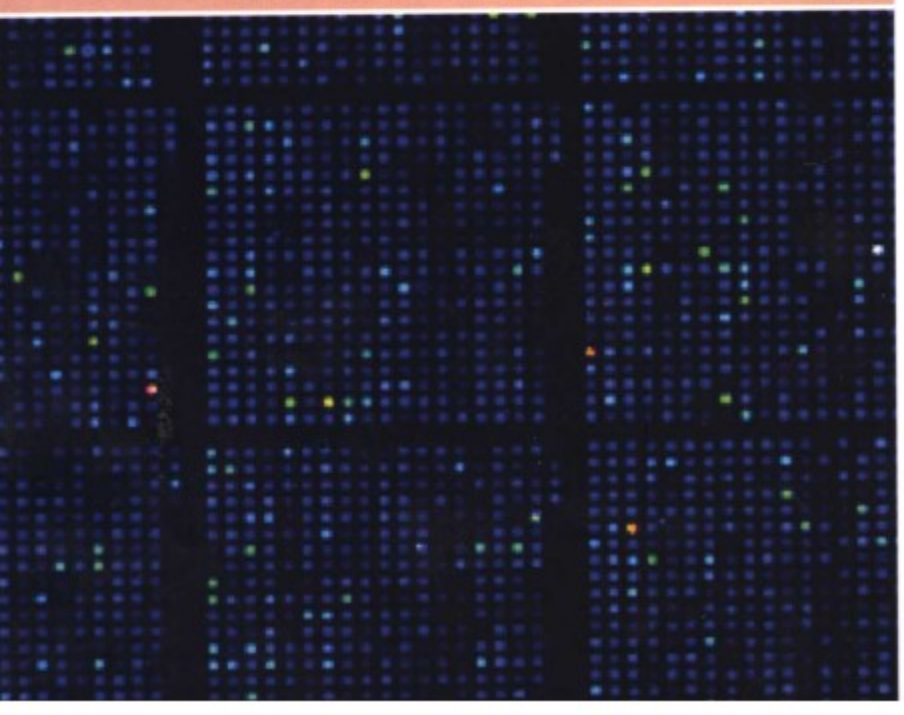
**版权所有 侵权必究**

物料号 22214-00

## 主要作者简介

朱玉贤，男，浙江省富阳市人。1989 年 12 月在美国康奈尔大学获得博士学位。先后在美国华盛顿大学和北京大学从事博士后研究。1997 年获得国家自然科学基金委“杰出青年科学基金”资助，2003 年成为“创新研究群体”主持人。现为北京大学教授、博士生导师，蛋白质工程及植物基因工程国家重点实验室主任。兼任教育部科学技术委员会生命科学学部委员，长江特聘教授，农业部“国家转基因生物安全委员会”委员，科技部“863”高技术计划“现代农业技术领域”专家组副组长。主持多项国家级科研项目。

除《现代分子生物学》外，还著有《分子生物学实验技术》，译有《PCR 传奇》和《细胞的起源》。



# 郑重声明

高等教育出版社依法对本书享有专有出版权。任何未经许可的复制、销售行为均违反《中华人民共和国著作权法》，其行为人将承担相应的民事责任和行政责任，构成犯罪的，将被依法追究刑事责任。为了维护市场秩序，保护读者的合法权益，避免读者误用盗版书造成不良后果，我社将配合行政执法部门和司法机关对违法犯罪的单位和个人给予严厉打击。社会各界人士如发现上述侵权行为，希望及时举报，本社将奖励举报有功人员。

**反盗版举报电话**：(010) 58581897/58581896/58581879

**传　　真**：(010) 82086060

**E - mail**：dd@hep.com.cn

**通信地址**：北京市西城区德外大街 4 号
高等教育出版社打击盗版办公室

**邮　　编**：100011

**购书请拨打电话**：(010)58581118

当你进入实验室时，要像脱去外衣那样放下你的想象力，因为实验操作中不能有一丁点儿的想象，否则，你对事物的观察就会受影响；当你翻开书本的时候，你又必须尽可能展开想象的“翅膀”，否则，你就不可能走在别人的前面。

# 第三版前言

我上小学的时候，有一句话特别风行，叫做“学好数理化，走遍天下都不怕”！当然，这话在“文革”的十年浩劫中与另一句同样广泛传播的俚语“龙生龙，凤生凤，老鼠生儿会打洞”一道受到严厉的批判，前者的罪名是“鼓励走白专道路”，后者的罪名则是“宣扬资产阶级血统论”。现在看来，“老鼠生儿会打洞”可能是生物学上一个永恒的命题，老鼠特有的DNA决定了它们的本能和生物学本质。一只老鼠，要是生而不能打洞，它的生存必将受到巨大的威胁，甚至不太可能活着离开父母的洞穴。

确实，直到20世纪后半叶，或者大致说直到整个20世纪结束以前，世界上的主要声音都是“科学推动社会进步”，都认为科学（特别是自然科学）是保证人类社会长期兴旺发达的最强大的动力。即使在20世纪30至40年代，中国处于军阀混战、外寇入侵、尸横遍野、民不聊生的凄惨境地，革命先驱们用以唤起普通民众的口号仍是“德先生（democracy）”和“赛先生（science）”（前者意为“民主”，后者意为“科学”）。

但是，对于地球——我们的母亲来说，科学其实更多地扮演了掠夺者、破坏者的角色。在过去的一二百年里，科学得到了突飞猛进的大发展，而我们的家园也在这一时期遭到了史无前例的大破坏。电能和动力机车的发现为化石燃料的大规模挖掘利用奠定了基础，滚滚向前的车轮、工厂的锅炉以及民居的取暖设施所释放的一氧化碳、二氧化碳和二氧化硫是导致全球性气候变暖的罪魁祸首；炸药以及稍后发现的核裂变、核聚变理论则成为战争狂人征服世界的工具。虽说没有科学做后盾，一棍一棒、一刀一石也能杀人，也能演变成战争，但肯定不至于像现在这样动辄多少万

吨级的炸药,炸你个遍地开花(!)没商量。此外,化肥、农药、塑料制品等数不胜数的科技成果在拓展人类生存空间、创造经济效益和生活便利的同时,也以令人难以置信的速度污染了环境,在不经意间剥夺了子孙后代生存繁衍的机会。

若干年前,我曾应邀参加美国某著名大学举办的一次学术年会,有学者做报告指出,对地球生物圈而言,一共只有三类物种:植物、微生物和其他。不幸得很,人类,骄傲的人类,只能归属到"其他"一族。当时倍觉愕然。现在想起来,这个划分虽然有以偏概全之嫌,但确实有深层次的含义。因为我们对地球基本没有正向的贡献,没有给予,只有索取,无穷无尽的索取。在人类历史的长河中,我们追求从"必然王国"向"自由王国"的进步,吹嘘如何"认识世界"、"改造世界",其目的却似乎永远只有一个:把地球改造得更适合人类居住和生活,顺便把废水和垃圾杂物或公开或悄悄地投入大江大海、山沟旮旯里,很少意识到应该为地球、为我们生存的大环境做些什么。

或许有人说,21 世纪是生命科学的世纪。因为,数学、物理学等实证科学(exact science)的进步不但带来了人类社会的物质文明,也带来了人口膨胀、环境污染、疾病猖獗、能源资源匮乏、生态平衡破坏等一系列前人所没有预料到的问题,而生命科学有可能成为解决这些问题的突破口。从 20 世纪下半叶开始,人类社会已经从崇尚实证科学逐步转变为注重分析科学(analytical science),包括生物学、计算机、信息科学、材料科学和环境、生态、资源保护等研究,注重人文科学,强调"和谐社会"(这里首先指人与地球大环境的和谐)。没有这个"和谐",人类社会的持续发展就可能成为泡影。

基于上述思考,本书在第 3 版修订中突出强调了分子生物学的实验技术和原理,在"疾病与人类健康"一章中增加了关于人禽流感和严重急性呼吸系统综合征(SARS)分子机制的讨论。根据学科发展的最新动态,对第 11 章(原第 10 章)"基因组与比较基因组学"做了较大规模的修改和充实。考虑到最近十多年来 RNA 研究持续升温,这一版增加了主要从事 RNA 研究的郑晓峰为第三作者,由她负责对第 2~4 章进行修改。李毅仍然负责与疾病和发育相关的第 9~10 章。全书其他章节的修改均由朱玉贤完成。

要真正学好、弄懂分子生物学这门重要的实验科学,除了在书本上下工夫,理解并掌握其基本规律之外,一定要善于跟踪和研究最新的科技文献,因为这些新发现、新进展往往对于诠释学习中碰到的疑难问题有着举一反三、画龙点睛的作用。此外,实验操作是掌握分子生物学精髓的主要途径,经常动手做实验或设计一些新的实验程序,往往能起到事半功倍的效果。希望读者、特别是广大青年学生,逐步拥有新的科学意识,在追求科学真理、实现远大抱负、"不怕"面对人生之旅的同时,学会"换位思考",用尊敬、崇尚和感恩的眼光看待大自然,让地球母亲在科学的呵护下生生不息!只有这样,科学才能不断得到发扬光大!

朱玉贤

2007 年 2 月 17 日写于燕园

# 第二版前言

分子生物学(Molecular Biology)是研究核酸等生物大分子的功能、形态结构特征及其重要性和规律性的科学,是人类从分子水平上真正揭开生物世界的奥秘,由被动地适应自然界转向主动地改造和重组自然界的基础学科。因为分子生物学所关心的既不是存在于生物体内的每一种分子,也不是每一种大分子,它所关心的主要是核酸在细胞生命过程中的作用,包括核酸本身的复制、保存以及基因(核酸大分子中最基本的功能单位)的表达与调控规律,所以,这门学科其实应该被叫做核酸生物学(Biology of Nucleic Acid)。当年的学者们为了图省事或者图新鲜,就发明了现在这个名词。尽管不十分贴切,却是言简意赅,容易被接受和传播,非常符合新兴大众传媒学的口味。一旦被炮制出来,这个学科名就不胫而走,深入人心,再也无法改变了。

追根寻源,分子生物学研究可能起源于德国。1869 年,Miescher 首次从莱茵河鲑鱼精子中提取了 DNA。1910 年,Kossel 第一次分离获得单核苷酸,揭开了核酸研究的序幕。可以把这短短的 130 多年大致分为前后两个时期。1940 年以前属于奠基阶段,通过孟德尔、摩尔根等人的努力,科学家开始认识生命遗传的分子基础。1940 年以后,在 Griffith 与 Avery 等人关于非致病菌发生遗传转化的实验、Hershey 与 Chase 证实 DNA 是遗传物质的实验基础上,Watson 和 Crick 提出了 DNA 反向平行双螺旋的结构模型,Jacob 和 Monod 则提出了阐明原核基因表达调控的操纵子学说,分子生物学迅速进入了大发展时期。时至今日,这一学科已经成为全世界生命科学乃至整个自然科学的主流,因为只有用分子手段才能研究和解答生命科学每一个分支中的根本性问题,使人

类掌握主动改造自然界的利剑,迎来生物学研究的新时代。

1991 年,本书的主要作者朱玉贤回国时,国内尚无系统性的分子生物学教材。那时,我国高校中还存在是否应为本科生开设“分子生物学”课程的争论。但是,世界上大多数发达国家的大学里已经有了这门课程。作者相信,有必要用现代生物学研究中最先进、最精彩的知识和技术来武装自己的学生,以帮助他们从分子水平上认识生命的本质,为毕业后参与尖端科学研究创造条件。为此,我们决定写一本能综合生命科学在分子水平上所取得的研究成果、所发现的基本规律与原理、所提供的新思维、新方法的教科书,为培养分子生物学专业人才并规范我国的研究生入学考试打基础。

1997 年,《现代分子生物学》第一次出版发行。北京大学根据原国家教委的文件精神举办了第一届全国“现代分子生物学教学研讨会”,包括北京大学、清华大学、复旦大学、中山大学、南京大学、南开大学和中国科技大学等全国重点院校在内的 30 多所高校派出 45 名骨干教师参加了研讨会。会议除交流现代分子生物学课程的教学方法、切磋教学经验之外,重点讨论了现代分子生物学的教学大纲和范畴,讨论了以《现代分子生物学》为基本教材进行本科生和研究生教学的可行性。到 2001 年,本书已连续重印 6 次,发行 4 万余册,成为国内许多高等院校本科生或研究生的主要教材。台湾艺轩图书出版社还在台湾和香港两地发行了繁体字版本,扩大了本教科书在世界上的影响。根据研讨会的精神,《现代分子生物学》第二版增加了果蝇体节发育、癌症和艾滋病的发生发展等内容,删去了第一版中有关植物分子生物学的大部分章节。新版本仍然分为十章,分别对染色体结构、DNA 的复制形式与特点、DNA 的转座、遗传密码的破译、蛋白质的合成和运转、基因表达调控的原理、癌症与癌基因活化、免疫缺损病毒(HIV)的分子机制等重大问题作了全面系统的分析,其中第三至九章以较大篇幅叙述了参与原核、真核细胞基因表达调控的各种元件,探讨了 DNA 甲基化、蛋白质磷酸化及各种不同环境因子对基因活性和基因功能的影响,第十章则讨论了基因组学与比较基因组学研究的最新成果。

虽然我们在这次修订过程中力求全面、完整地反映分子生物学各领域的最新进展,但面对浩如烟海的文献资料,加上作者本人的水平和能力有限,归纳成书后疏漏及错误之处肯定不少,殷切希望读者提出批评指正。我们将在教学研究的过程中不断修正并完善本教材,使之成为广大青年学生成长道路上的伴侣和助手。

朱玉贤

2002 年 2 月 15 日写于北大蓝旗营

# 第一版序言

本世纪初以来，生命科学所取得的巨大成就和进步，不但使生物学这门古老的学科焕发了青春，也使它在自然科学中的地位发生了革命性的变化。同时，生物学革命也为物理学、数学、化学、信息科学、材料与工程科学注入了大量新鲜血液，提出了数不胜数的新问题、新概念和新思路。它在各个科学之间广泛渗透，相互交叉，相互作用，极大地推动了科学的发展。种种迹象表明，生物学已经成为带头学科之一，领导着世界科技大军走向2000年。

可以说，任何自然科学研究都没有比人类彻底认识自己，了解自己，找出解决自身所面临的人口膨胀、粮食短缺、环境污染、疾病猖獗、能源资源匮乏、生态平衡被破坏及生物物种消亡等一系列问题更为重要、更为迫切，因而更具有吸引力。加强生物学研究，从分子水平、细胞水平、个体和群体水平等不同层次深入探索生命与自然的奥秘，全面改造和改良我们的生存环境与生存质量，正在日益成为世界上千百万有识之士的共同愿望，成为历史的潮流。分子生物学作为生物学科最新兴、最具活力的科学，在推动我国科学事业的发展、推动生物工程产业的崛起、推动国民经济持续高速发展等方面均有着举足轻重的影响。

落后就要挨打，就要受人宰割。没有强大的分子生物学基础研究，我们就不可能在生物工程这个21世纪的龙头产业中占有一席之地，就不可能与世界列强平等对话。知识就是力量，就是财富！谁拥有了先进的科学技术，谁就拥有了在世界上立于不败之地的法宝。衷心地祝愿我国的科学教育事业乘改革开放之强劲东

风，跃马扬鞭自奋蹄，以更快的速度向更高更新的目标前进！衷心地希望《现代分子生物学》能成为广大青年学生和科技工作者迈向分子生物学殿堂的钥匙和助手！

1995 年 10 月 5 日

# 目　录


1　绪论 …… 1

1.1　引言 …… 1

1.1.1　创世说与进化论 …… 1

1.1.2　细胞学说 …… 3

1.1.3　经典生物化学和遗传学 …… 4

1.1.4　DNA 的发现 …… 6

1.2　分子生物学简史 …… 8

1.3　分子生物学的主要研究内容 …… 11

1.3.1　DNA 重组技术 …… 12

1.3.2　基因表达调控研究 …… 13

1.3.3　生物大分子的结构功能研究——
结构分子生物学 …… 14

1.3.4　基因组、功能基因组与生物信息学研究 …… 14

1.4　展望 …… 15

2　染色体与 DNA …… 18

2.1　染色体 …… 19

2.1.1　染色体概述 …… 19

2.1.2　真核细胞染色体的组成 …… 22

2.1.3　原核生物基因组 …… 30

2.2　DNA 的结构 …… 31

2.2.1　DNA 的一级结构 …… 32

2.2.2　DNA 的二级结构 …… 33

2.2.3 DNA 的高级结构 …… 36
2.3 DNA 的复制 …… 37
2.3.1 DNA 的半保留复制 …… 38
2.3.2 复制的起点、方向和速度 …… 38
2.3.3 复制的几种主要方式 …… 40
2.4 原核生物和真核生物 DNA 复制的特点 …… 44
2.4.1 原核生物 DNA 复制的特点 …… 44
2.4.2 真核生物 DNA 复制的特点 …… 48
2.4.3 DNA 复制的调控 …… 50
2.5 DNA 的修复 …… 51
2.5.1 错配修复 …… 52
2.5.2 切除修复 …… 53
2.5.3 重组修复 …… 54
2.5.4 DNA 的直接修复 …… 55
2.5.5 SOS 反应 …… 55
2.6 DNA 的转座 …… 56
2.6.1 转座子的分类和结构特征 …… 57
2.6.2 真核生物中的转座子 …… 59
2.6.3 转座作用的机制 …… 62
2.6.4 转座作用的遗传学效应 …… 63

3 生物信息的传递(上)——从 DNA 到 RNA …… 66

3.1 RNA 转录的基本过程 …… 67
3.1.1 模板识别 …… 67
3.1.2 转录起始 …… 68
3.1.3 转录延伸 …… 69
3.1.4 转录终止 …… 69
3.2 转录机器的主要成分 …… 69
3.2.1 RNA 聚合酶 …… 69
3.2.2 转录复合物 …… 73
3.3 启动子与转录起始 …… 75
3.3.1 启动子区的基本结构 …… 75
3.3.2 启动子区的识别 …… 77
3.3.3 RNA 聚合酶与启动子区的结合 …… 78
3.3.4 -10 区与-35 区的最佳间距 …… 78
3.3.5 增强子及其功能 …… 78
3.3.6 真核生物启动子对转录的影响 …… 80
3.3.7 转录的抑制 …… 82
3.4 原核与真核生物 mRNA 的特征比较 …… 83
3.4.1 原核生物 mRNA 的特征 …… 84
3.4.2 真核生物 mRNA 的特征 …… 87
3.5 终止和抗终止 …… 90
3.5.1 不依赖于 ρ 因子的终止 …… 90
3.5.2 依赖于 ρ 因子的终止 …… 90
3.5.3 抗终止 …… 91
3.6 内含子的剪接、编辑、再编码及化学修饰 …… 93
3.6.1 RNA 中的内含子 …… 93
3.6.2 mRNA 的剪接 …… 96
3.6.3 RNA 的编辑、再编码及化学修饰 …… 99
3.6.4 核酶 …… 102
3.6.5 RNA 在生物进化中的地位 …… 105

4 生物信息的传递(下)——从 mRNA 到蛋白质 …… 107

4.1 遗传密码——三联子 …… 108
4.1.1 三联子密码及其破译 …… 108
4.1.2 遗传密码的性质 …… 111
4.2 tRNA …… 115
4.2.1 tRNA 的 L 形三级结构 …… 117
4.2.2 tRNA 的功能 …… 118
4.2.3 tRNA 的种类 …… 119
4.2.4 氨酰- tRNA 合成酶 …… 120
4.3 核糖体 …… 121
4.3.1 核糖体的结构 …… 123
4.3.2 核糖体的功能 …… 126
4.4 蛋白质合成的生物学机制 …… 126
4.4.1 氨基酸的活化 …… 127
4.4.2 翻译的起始 …… 129
4.4.3 肽链的延伸 …… 132
4.4.4 肽链的终止 …… 136
4.4.5 蛋白质前体的加工 …… 136

4.4.6 蛋白质的折叠 …… 140
4.4.7 蛋白质合成的抑制剂 …… 140
4.5 蛋白质转运机制 …… 142
4.5.1 翻译–转运同步机制 …… 144
4.5.2 翻译后转运机制 …… 146
4.5.3 核定位蛋白的转运机制 …… 150
4.5.4 蛋白质的降解 …… 151

5 分子生物学研究法(上)——DNA、RNA 及蛋白质操作技术 …… 155

5.1 重组 DNA 技术回顾 …… 155
5.2 DNA 基本操作技术 …… 161
5.2.1 核酸凝胶电泳技术 …… 161
5.2.2 细菌转化与目标 DNA 分子的增殖 …… 163
5.2.3 聚合酶链反应技术 …… 165
5.2.4 实时定量 PCR …… 167
5.2.5 基因组 DNA 文库构建 …… 170
5.3 RNA 基本操作技术 …… 172
5.3.1 总 RNA 的提取 …… 172
5.3.2 mRNA 的纯化 …… 173
5.3.3 cDNA 的合成 …… 174
5.3.4 cDNA 文库的构建 …… 175
5.3.5 基因文库的筛选 …… 176
5.4 SNP 的理论与应用 …… 178
5.4.1 SNP 概述 …… 178
5.4.2 SNP 的检测技术 …… 179
5.4.3 SNP 的应用 …… 182
5.5 基因克隆技术 …… 182
5.5.1 RACE 技术 …… 183
5.5.2 应用 cDNA 差示分析法克隆基因 …… 184
5.5.3 Gateway 大规模克隆技术 …… 185
5.5.4 基因的图位克隆法 …… 185
5.6 蛋白质组与蛋白质组学技术 …… 187
5.6.1 双向电泳技术 …… 188
5.6.2 蛋白质印迹法 …… 189
5.6.3 蛋白质的质谱分析技术 …… 191

6 分子生物学研究法(下)——基因功能研究技术 …… 193

6.1 基因表达研究技术 …… 193
6.1.1 基因表达系列分析技术 …… 193
6.1.2 RNA 的选择性剪接技术 …… 195
6.1.3 原位杂交技术 …… 197
6.1.4 基因定点突变技术 …… 197
6.2 基因敲除技术 …… 200
6.2.1 基本原理 …… 200
6.2.2 高等动物基因敲除技术 …… 201
6.2.3 植物基因敲除技术 …… 205
6.3 蛋白质及 RNA 相互作用技术 …… 206
6.3.1 酵母单杂交系统 …… 206
6.3.2 酵母双杂交系统 …… 207
6.3.3 体外蛋白质相互作用技术 …… 208
6.3.4 细胞内蛋白质相互作用研究——荧光共振能量转移法 …… 212
6.3.5 RNAi 技术及其应用 …… 212
6.4 基因芯片及数据分析 …… 214
6.4.1 基因芯片技术原理 …… 214
6.4.2 基因芯片的点制过程 …… 215
6.4.3 基因芯片数据分析 …… 217
6.5 利用酵母鉴定靶基因功能 …… 218
6.5.1 酵母的遗传学和分子生物学简介 …… 218
6.5.2 酵母基因转化与性状互补 …… 220
6.5.3 外源基因在酵母中的功能鉴定 …… 222
6.6 其他分子生物学技术 …… 223
6.6.1 凝胶滞缓实验 …… 223
6.6.2 噬菌体展示技术 …… 225
6.6.3 蛋白质磷酸化分析技术 …… 226

7 基因的表达与调控(上)——原核基因表达调控模式 …… 230

7.1 原核基因表达调控总论 …… 231
7.1.1 原核基因调控分类 …… 232

7.1.2 原核基因调控的主要特点 …… 235
7.1.3 弱化子对基因活性的影响 …… 235
7.1.4 降解物对基因活性的调节 …… 236
7.1.5 细菌的应急反应 …… 236
**7.2 乳糖操纵子与负控诱导系统** …… 237
7.2.1 酶的诱导——*lac* 体系受调控的证据 …… 238
7.2.2 操纵子模型及其影响因子 …… 239
7.2.3 *lac* 操纵子 DNA 的调控区域——P 区和 O 区 …… 244
7.2.4 *lac* 操纵子中的其他问题 …… 245
**7.3 色氨酸操纵子与负控阻遏系统** …… 246
7.3.1 *trp* 操纵子的阻遏系统 …… 246
7.3.2 弱化子与前导肽 …… 246
7.3.3 *trp* 操纵子弱化机制的实验依据 …… 251
7.3.4 *trp* 操纵子的其他调控机制 …… 253
**7.4 其他操纵子** …… 255
7.4.1 半乳糖操纵子 …… 255
7.4.2 阿拉伯糖操纵子 …… 257
7.4.3 阻遏蛋白 LexA 的降解与细菌中的 SOS 应答 …… 260
7.4.4 二组分调控系统和信号转导 …… 261
7.4.5 多启动子调控的操纵子 …… 262
**7.5 固氮基因调控** …… 263
7.5.1 根瘤的产生 …… 263
7.5.2 固氮酶 …… 264
7.5.3 与固氮有关的基因及其调控 …… 265
**7.6 转录水平上的其他调控方式** …… 269
7.6.1 $\sigma$ 因子的调节作用 …… 269
7.6.2 组蛋白类似蛋白的调节作用 …… 270
7.6.3 转录调控因子的作用 …… 270
7.6.4 抗终止因子的调节作用 …… 271
**7.7 转录后调控** …… 272
7.7.1 mRNA 自身结构元件对翻译起始的调节 …… 272
7.7.2 mRNA 稳定性对转录水平的影响 …… 273
7.7.3 调节蛋白的调控作用 …… 274
7.7.4 反义 RNA 的调节作用 …… 274
7.7.5 稀有密码子对翻译的影响 …… 275
7.7.6 重叠基因对翻译的影响 …… 276
7.7.7 翻译的阻遏 …… 277
7.7.8 魔斑核苷酸水平对翻译的影响 …… 277

**8 基因的表达与调控(下)——真核基因表达调控一般规律** …… 280
**8.1 真核生物的基因结构与转录活性** …… 282
8.1.1 基因家族 …… 282
8.1.2 真核基因的断裂结构 …… 286
8.1.3 真核生物 DNA 水平上的基因表达调控 …… 288
8.1.4 DNA 甲基化与基因活性的调控 …… 292
**8.2 真核基因转录机器的主要组成** …… 296
8.2.1 真核基因的转录 …… 297
8.2.2 增强子及其对转录的影响 …… 298
8.2.3 反式作用因子 …… 299
**8.3 蛋白质磷酸化对基因转录的调控** …… 307
8.3.1 受 cAMP 水平调控的 A 激酶 …… 310
8.3.2 C 激酶与 $PIP_2$、$IP_3$ 和 DAG …… 311
8.3.3 CaM 激酶及 MAP 激酶 …… 312
8.3.4 酪氨酸激酶途径 …… 313
8.3.5 蛋白质磷酸化与细胞分裂调控 …… 315
**8.4 蛋白质乙酰化对基因表达的影响** …… 316
8.4.1 组蛋白的乙酰化及去乙酰化 …… 316
8.4.2 组蛋白乙酰化及去乙酰化对

基因表达的影响 …… 320
8.4.3 p53 乙酰化对转录活性的影响 …… 320
8.5 激素与热激蛋白对基因表达的影响 …… 322
8.5.1 激素对靶基因的影响 …… 322
8.5.2 热休克蛋白诱导的基因表达 …… 325
8.6 其他水平上的表达调控 …… 329
8.6.1 RNA 的加工成熟 …… 329
8.6.2 翻译水平的调控 …… 333
9 疾病与人类健康 …… 338
9.1 肿瘤与癌症 …… 338
9.1.1 反转录病毒致癌基因 …… 339
9.1.2 原癌基因(细胞转化基因)产物及其分类 …… 343
9.1.3 原癌基因的表达调控 …… 346
9.1.4 基因相互作用与癌基因表达 …… 349
9.2 人类免疫缺陷病毒——HIV …… 351
9.2.1 HIV 病毒颗粒的形态结构及传播 …… 352
9.2.2 HIV 基因组及其编码的蛋白 …… 352
9.2.3 HIV 的复制 …… 355
9.2.4 HIV 基因表达调控 …… 356
9.2.5 HIV 的感染及致病机制 …… 362
9.2.6 艾滋病的治疗及预防 …… 363
9.3 乙型肝炎病毒——HBV …… 364
9.3.1 HBV 的分类及病毒粒子结构 …… 364
9.3.2 HBV 基因组及其编码的蛋白 …… 365
9.3.3 HBV 的复制 …… 368
9.4 人禽流感的分子生物学机制 …… 369
9.4.1 人禽流感病毒特点及分型 …… 370
9.4.2 禽流感病毒进入细胞及转录与复制 …… 371
9.4.3 禽流感病毒感染人类的机制 …… 372
9.5 严重急性呼吸系统综合征-SARS 的分子机制 …… 373
9.5.1 SARS-CoV 的结构与分类 …… 373
9.5.2 SARS-CoV 的基因组结构 …… 374
9.5.3 SARS-CoV 的侵染过程 …… 375
9.5.4 SARS-CoV 的起源及变异 …… 376
9.6 基因治疗 …… 377
9.6.1 基因治疗的主要途径 …… 377
9.6.2 基因治疗中的病毒载体 …… 378
9.6.3 基因治疗中的非病毒载体 …… 381
10 基因与发育 …… 384
10.1 免疫系统发育及免疫球蛋白基因表达 …… 384
10.1.1 脊椎动物免疫系统 …… 384
10.1.2 B-淋巴细胞的发育和分化 …… 385
10.1.3 T-淋巴细胞发育及分化 …… 388
10.1.4 免疫球蛋白的结构 …… 389
10.1.5 免疫球蛋白基因结构 …… 390
10.1.6 Ig 基因重排与 DNA 的多样性 …… 392
10.1.7 免疫球蛋白基因表达 …… 396
10.1.8 主要组织相容复合体的表达调控 …… 397
10.2 果蝇的发育与调控 …… 400
10.2.1 卵子发育 …… 400
10.2.2 胚胎发育 …… 401
10.2.3 果蝇的末端系统及背腹极性基因与发育调控 …… 404
10.2.4 影响果蝇体节发育的基因 …… 405
10.2.5 果蝇的同源域基因 …… 406
10.3 高等植物花发育的基因调控 …… 406
10.3.1 植物的花器官结构 …… 407
10.3.2 调控花器官发育的主要基因 …… 407
10.3.3 花器官发育的“ABC”模型 …… 409

10.3.4 参与诱导开花信号产生、传递和感知的基因 …… 410

**11 基因组与比较基因组学 …… 416**

**11.1 高通量 DNA 序列分析技术 …… 417**

11.1.1 DNA 序列测定的基本原理 …… 417

11.1.2 基因组 DNA 大片段文库的构建 …… 418

11.1.3 鸟枪法序列测定技术及其改良 …… 419

11.1.4 大规模 DNA 序列拼接 …… 422

**11.2 人类基因组计划 …… 423**

11.2.1 人类基因组计划的科学意义 …… 424

11.2.2 遗传图 …… 425

11.2.3 物理图 …… 431

11.2.4 转录图 …… 432

11.2.5 全序列图 …… 434

**11.3 其他基因组 …… 436**

11.3.1 大肠杆菌基因组 …… 438

11.3.2 酵母基因组 …… 439

11.3.3 拟南芥基因组 …… 440

11.3.4 水稻基因组 …… 440

11.3.5 家蚕基因组 …… 441

11.3.6 鸡基因组 …… 442

**11.4 比较基因组学及相关研究 …… 443**

11.4.1 基因组数据的挖掘与分析 …… 444

11.4.2 基因组数据的比较研究 …… 445

**名词解释 …… 448**

**主要参考书目 …… 464**

**第三版后记 …… 465**

**索引 …… 466**

# 1 绪　论


1.1 引言
1.2 分子生物学简史
1.3 分子生物学的主要研究内容
1.4 展望


## 1.1 引言

现代生物学研究的目标是要在分子水平上阐明细胞活动的规律，揭示生命的本质。从20世纪40年代开始，无数生命科学家用他们的智慧和汗水，赢得了20世纪自然科学最伟大的革命——揭开生物遗传的谜底。随着DNA的结构与功能、RNA在蛋白质合成中的作用、蛋白质的结构与功能、遗传密码及基因表达调控的本质等被相继阐明，人类开始了从生物学的必然王国向自由王国的过渡。分子水平的生物学研究，正在越来越多地影响着各个传统生物科学领域，如组织学、细胞学、解剖学、胚胎学、遗传学、生理学和进化论。本书将尽可能系统性地提供有关现代分子生物学各分支的基本理论和主要实验依据，本章首先介绍现代生物学的历史背景和主要人物，并简单讨论人类对遗传的最基本单位——基因化学本质的认识过程。

### 1.1.1 创世说与进化论

多少年来，人们常常会反复提出3个与生命和一切生物学现象有关的问题：

① 生命是怎样起源的？

② 为什么“有其父必有其子”？

③ 动、植物个体是怎样从一个受精卵发育而来的？

直到19世纪初，这些问题还只能从宗教或迷信的角度上进行回答。西方人一直相信基督教的宣传，相信上帝先创造了

花草树木、世间万物，后来又创造了男人亚当，再从亚当身上抽出一根肋骨，这就成了女人夏娃。亚当、夏娃婚配繁衍产生了人类。1859 年，伟大的英国生物学家达尔文(Charles Darwin)发表了著名的《物种起源》，确立了进化论的概念。正是达尔文的生物进化学说，打破了西方社会上帝造人的传统观念，改变了社会对人类在整个世界中的地位的看法，极大地推动了人类思想的发展。

达尔文从小热爱大自然，喜欢采集动、植物标本。他 16 岁到爱丁堡大学学习，参加了青年人组织的普林尼学术活动，共同研讨拉马克的进化学说。拉马克虽然不信“上帝创造一切”的“创世说”，却又拿不出令人信服的证据来。这些讨论使达尔文的思想陷于矛盾和斗争之中，他决心深入大自然去寻找答案。

9 年后，达尔文以自然科学家的身份，参加了历时 5 年的贝格尔号军舰环球旅行，历尽了千辛万苦，即使在晕船、饥渴、病痛和死亡的威胁下，他也坚持工作。他观察过火山，经历过地震，见到了形形色色稀奇古怪的动物和植物。达尔文采集了大量动、植物标本和化石并细心地进行比较、鉴别和研究，提出并解答了一系列学术问题，如相似的动物为什么居住在千里之外的不同地区？同一个小岛上为什么聚集着许多不同的动物？低等动物与高等动物有些什么样的联系？人是如何产生的……。他认为，大陆自古以来发生过许多次巨变，如冰川时期等，所以不可能存在亘古不变的动、植物。

从贝格尔号回到英国以后，他发表了一系列论文，逐步阐述了生物进化的观点。在《物种起源》这部划时代的科学巨著中，他用大量事实证明“物竞天择，适者生存”的进化论思想。他认为世界上的一切生物都是可变的，并预言从低级到高级的变化过程中必定有过渡物种存在。他指出物种的变异是由于大自然的环境和生物群体的生存竞争造成的，彻底否定了上帝创造万物的旧思想，推翻了物种不变的神话，使生物学真正迈入实证自然科学的行列。

通过记载不同动、植物的地理分布，研究近亲种族的解剖学、形态学的相似性和变异率，达尔文第一个认识到生物世界的不连续性。他还发现，当记录研究跨越一个较长的历史时期，主要存在物种会有很大的变化。他提出许多环境因素，如大地变迁、特定区域内的温度、降雨量变化及气候条件改变，都会以“自然选择压力”的形式，在生物体的世代遗传中体现出来。正是在这种“自然选择压力”之下，新物种才不断诞生，旧的、与环境不再相容的物种也不断消亡。他在书中这样写道：对于每一个动、植物种群来说，因为总是有大大多于可能生存下来的个体出生，所以为生存而斗争是长期的、永久的。如果某些个体偶然获得了于自身有利的变异就会在生与死的斗争中占同类的上风，从而生存下来。根据遗传学原理，任何生存下来的个体都倾向于扩增其经过修饰的新性状，以保持生存优势。

达尔文关于生物进化的学说及其唯物主义的物种起源理论，是生物科学史上最伟大的创举之一，具有不可磨灭的贡献。为了纪念这位生物科学大师，人们把进化论称为“达尔文学说”。

### 1.1.2 细胞学说

早期生物科学家的另一大贡献是提出了细胞理论(cell theory)。17 世纪末，荷兰籍显微镜学家 Leeuwenhoek 成功制造了世界第一台光学显微镜。通过这一装置，他看到了一系列肉眼看不到而又使人迷惑不解的微小生物，他将这些小生命称为“微动物”(animalcule)。若干年后，人们才知道它们是单细胞生物。

Leeuwenhoek 出身贫寒，16 岁便失学当了学徒。在好奇心驱使下，他把工余时间都用来研究、磨制、装配玻璃透镜。开始，他用自己磨制的透镜观察蜜蜂蜇人的“针”，看蚊子叮人的嘴，以及小甲虫的腿等。随着制镜手艺的不断提高，他制成了能放大 200 倍的显微镜，他不断公布自己的观察结果，并将新发现报告给当时世界最权威的科学管理机构——英国皇家学会。他第一个观察到狗和人的精子，发现了酵母菌，描述了红细胞等。为了表彰和鼓励 Leeuwenhoek 的研究工作，英国皇家学会吸收他为会员。一个小学徒终于成为受人尊敬的科学家。

1702 年，Leeuwenhoek 在观察轮虫时，偶然发现雨水中有微生物。这些生物是怎么来的呢？为了解开这个谜，他做了一个实验：收集开始下雨时的雨水来观察，里面并没有微生物。到了第四天再观察，就有许多微生物出现在水中。Leeuwenhoek 因此得出了一个结论：风能将空气灰尘中的微生物带入水中。以后经过对昆虫、海贝和鳝鱼等的细心研究，他进一步断定：微生物不是由泥沙尘埃产生的，而是和动物一样，有完整的生活史。这一有趣的发现使 Leeuwenhoek 更为出名。

大约与 Leeuwenhoek 同时代的 Hooke，第一次用“细胞”这个概念来形容组成软木的最基本单元。虽然直到 19 世纪中叶，这一概念才正式被科学界所接受，但它对生物学的贡献是不可估量的。随着显微技术、组织保存技术和超薄切片技术的不断发展，科学家发现动、植物组织都是由细胞所组成，而且细胞是可以分裂的，每一个细胞都是或曾经是一个单独的活的实体，包含有生命的全部特征。

动、植物的基本单元是细胞，这是 19 世纪三大发现之一的细胞学说的核心。建立这一学说的是德国植物学家 Schleiden 和动物学家 Schwann。Schleiden 出生于汉堡，22 岁就获得了法学博士学位，但他并不喜欢当律师。28 岁时，他到哥廷根和柏林学习植物学和医学，35 岁时获得医学和哲学博士学位。Schwann 是首饰匠的儿子，16 岁高中毕业后，没有按照父母的意愿学神学，而毅然去柏林学医。24 岁获得博士学位，在柏林解剖博物馆工作时结识了 Schleiden。他俩虽然个性、经历迥然不同，但共同的志趣和真诚的情感促成了他们多年的合作。Schleiden 研究被子植物的胚囊，Schwann 研究蛙类的胚胎组织。相同的研究方向，相似的研究方法，使他们取得了一致见解，共同创立了生物科学的基础理论——细胞学说。

1847 年，Schwann 在描述动物组织时这样写道：所有组织的最基本单元是形状非常相似而又高度分化的细胞。可以认为，细胞的发生和形成是生物学界普遍和永久的规律。

从此,细胞学说开始广为传播。越来越多的科学家发现,每一个动、植物个体实际上是千千万万个生命单元的总和,而这些微小单元——细胞,包含了所有的生命信息。细胞学说对生物科学最重要的贡献在于:因为单个细胞生长分裂,组织、器官和个体的生命现象实际上是细胞活动的总和,所以细胞可以而且应该成为生物学研究的首要对象。今天的细胞学和分子细胞学就是在这个基础上发展起来的。

### 1.1.3 经典生物化学和遗传学

进化论和细胞学说相结合,产生了作为主要实验科学之一的现代生物学,而以研究动、植物遗传变异规律为目标的遗传学和以分离纯化、鉴定细胞内含物质为目标的生物化学则是这一学科的两大支柱。早在19世纪中叶,人们就发现动物和植物细胞的提取液中主要是一些能受热或酸变性形成纤维状沉淀的物质。这些物质包含有大体相等的物质的量浓度的碳、氢、氧和氮。科学家将这些物质命名为蛋白质。生物化学家Buchner第一个实现了用酵母无细胞提取液和葡萄糖进行氧化反应,生成乙醇,证明化学物质转换并不需要完整的细胞而仅仅需要细胞中的某些成分。蛋白质是生活细胞中所有化学反应的执行者和催化剂。

生物化学从一开始就执行着双重使命。首先,分析细胞的组成成分;其次,弄清楚这些物质与细胞内生命现象的联系。19世纪中叶到20世纪初,是早期生物化学的大发展阶段,组成蛋白质的20种基本氨基酸被相继发现(最晚分离的是苏氨酸,1935年),著名生物化学家Fisher还论证了连接相邻氨基酸的"肽键"的形成。细胞的其他部分,如脂质、糖类和核酸也相继在那一阶段被科学家所认识和部分纯化。当时,科学家还无法解释细胞内最重要的生命活动,即细胞成分是如何世代相传的。

奥地利大科学家、经典遗传学的创始人孟德尔(Gregor Mendel)发现并提出遗传学定律的故事像是不朽的神话,在生物学界被广泛传诵。

孟德尔从小爱好园艺,虽然因为家境贫寒,没有念完大学就当了修道士,但他却矢志不渝地钻研科学。开始时,他对"种瓜得瓜,种豆得豆"的生物遗传现象感到好奇和困惑,就在修道院里种了许多花木,还挑选了20多种大小不同,形状、颜色各异的食用豌豆,反复进行杂交、自交等试验,并做了详细的记载。从1857年到1864年的7年间,孟德尔选择了7对差异明显的简单性状,对豌豆的生长进行了仔细的观察。例如,他用产生圆形种子的豌豆同产生皱皮种子的植株杂交,得到几百粒全是圆形的杂交子一代($F_1$)种子。第二年,他种植了253粒$F_1$圆形种子并进行自交,得到7 324粒$F_2$种子,他发现有5 474粒是圆形的,1 850粒是皱皮的,用统计学方法计算出圆皱比为3:1。

他还进行了具有2个对立性状的豌豆品系之间的双因子杂交试验。他发现当选用产生黄色圆形种子的豌豆品系同产生绿色皱皮种子的豌豆品系进行杂交时,所产生的$F_1$种子全是黄色圆形的,但在自交产生的$F_2$代576粒种子中,不但出现了2种亲本类型,而且还出现了2种新的重组类型,其中黄色圆形315粒,黄色皱皮121粒,绿色圆形108粒,绿色皱皮32粒。这4种类型的比例接近于9:3:3:1。

根据以上现象,孟德尔总结出生物遗传的两条基本规律:

第一,当两种不同植物杂交时,它们的下一代可能与亲本之一完全相同。他把这一现象称为统一律。根据自己长期的实验结果,孟德尔认为,生物的每一种性状都是由遗传因子控制的。这些因子可以从亲代到子代,代代相传。在体细胞内,遗传因子是成对存在的,其中一个来自父本,一个来自母本。在形成配子时成对的遗传因子彼此分开,单独存在。他还认为,有些遗传因子以显性(dominant)形式存在的,即能在任何杂种一代得到表达;而有些因子呈隐性(recessive)状态,只有当父、母本同时含有这一因子时,才得到表现。

第二,将不同植物品种杂交后的 $F_1$ 代种子再进行杂交或自交时,下一代就会按照一定的比例发生分离,因而具有不同的形式。他把这一现象称为分离规律。如把红花和白花杂交,得到的 8 株 $F_1$ 代植株全部为灰色花。灰色花互相交配结果得到 2 株红花、2 株白花、4 株灰色花。这一代的白色花互相交配,将永远得到白色花,红色花互相交配得到的永远是红色花。但这一代的灰色花互相交配,其结果就像上一代那样,仍旧是 2 红、2 白、4 灰。所有这些花朵,都按照孟德尔的分离规律依次遗传下去。分离规律对每一代动、植物都适用。

由于孟德尔的研究方法和结论都远远超过了当时的科学认识水平,他的这些天才的科学发现和见解,并没有立即引起生物学界的注意。从 1865 年他发表《植物杂交试验》一文到 1884 年逝世,欧美各国科学界几乎无人理睬他的巨大贡献。直到 1900 年,他的理论才被重新发现并得到普遍应用,并逐渐被公认为经典遗传学的奠基人。

虽然孟德尔早在 1861 年就通过豌豆杂交试验揭示了遗传的物质性(adiscrete unit governing inherited characteristics),直到 1909 年丹麦科学家 Wilhelm Ludvig Johannsen(1857—1927)才首次用“基因(gene)”这个词代表遗传学的最基本单位。美国人托马斯·摩尔根(Thomas Hunt Morgan,1866—1945)则是第一个用实验证明“基因”学说的科学家。1910 年,Morgan 和他的助手们发现了第一只白眼雄果蝇。因为正常情况下,果蝇都是红眼的,称为野生型种质,所以,他们将白眼果蝇称为突变型种质。到1915 年,他们一共找到 85 种果蝇的突变型。这些突变型果蝇跟野生型相比,在翅长、体色、刚毛形状、复眼数目等性状上都有差别。有了这些突变型,就能更广泛地进行杂交试验,从而更加深入地进行遗传机制的探讨。Morgan 将白眼雄果蝇与红眼雌果蝇交配,所产生的 $F_1$ 代不论雌雄,全为红眼果蝇(孟德尔的统一规律!)。让这些 $F_1$ 果蝇互相交配所产生的 $F_2$ 代有红眼,也有白眼,但有趣的是所有的白眼果蝇都是雄性的,说明白眼性状与性别有联系,这一点与孟德尔的遗传性状独立分离规律是背道而驰的(现在我们已经知道,当所研究的两个基因位于同一条染色体上而又距离较近时,Morgan 的连锁遗传规律起主导作用;而当所研究的两个基因位于不同染色体上时,孟德尔的独立分离规律起主导作用)。

为了解释这些现象,有必要对果蝇的染色体做一下简单介绍。果蝇只有 4 对染色体。在雌果蝇中,有 1 对很小呈颗粒状的染色体,2 对呈“V”形的染色体,另有 1 对呈棒状、称为 XX 的染色体。在雄果蝇体内,前 3 对同雌果蝇完全相同,但缺

少棒状的 XX 染色体,而由一条棒状的 X 染色体和一条呈"J"形的 Y 染色体所取代,人们称这一对为 XY 染色体。Morgan 当时就知道性染色体的存在,因此他推想,白眼这一隐性基因($w$)是位于 X 染色体上,而在 Y 染色体上没有它的等位基因。他让 $F_1$ 红眼雌果蝇($Ww$)与白眼雄果蝇亲本($wY$)回交,结果产生的后代果蝇中有 1/4 是红眼雌果蝇,1/4 是白眼雄果蝇。这个实验证明,白眼隐性突变基因($w$)确实位于 X 染色体上。这一现象被称为遗传性状的连锁定律,又称连锁遗传。

Morgan 和他的助手们第一次将代表某一特定性状的基因,同某一特定的染色体联系起来,使科学界普遍认识了染色体的重要性并接受了孟德尔的遗传学原理。Morgan 特别指出:种质必须由某些独立的要素组成,我们把这些要素称为遗传因子,或者更简单地称为基因。

### 1.1.4 DNA 的发现

尽管由于 Morgan 及其学派的出色工作,基因学说得到了普遍承认,但是,直到 1953 年 Watson 和 Crick 提出 DNA 双螺旋模型之前,人们对于基因的理解仍然是抽象的、概念化的,缺乏准确的物质内容。那时的遗传学家,不但没有探明基因的结构特征,而且也不能解释位于细胞核中的染色体和基因是怎样控制显然发生在细胞质中的各种生物化学过程,不能解释基因是怎样在细胞繁殖过程中准确地复制和遗传的。

早在 1928 年英国科学家 Griffith 等人就发现,肺炎链球菌使小鼠死亡的原因是引起肺炎。细菌的毒性(致病力)是由细胞表面荚膜中的多糖所决定的。具有光滑外表的 S 型肺炎链球菌因为带有荚膜多糖而能使小鼠发病,具有粗糙外表的 R 型细菌因为没有荚膜多糖而失去致病力(荚膜多糖能保护细菌免受动物白细胞的攻击)。

首先用实验证明基因就是 DNA 分子的是美国著名的微生物学家 Avery。他和他的同事们首先将光滑型致病菌(S 型)烧煮杀灭活性以后再侵染小鼠,发现这些死细菌自然丧失了致病能力(图 1-1)。再用活的粗糙型细菌(R 型)来侵染小鼠,也不能使之发病,因为粗糙型细菌天然无致病力。然而,当他们将经烧煮杀死的 S 型细菌和活的 R 型细菌混合再感染小鼠时,奇迹发生了! 实验小鼠每次都可怜巴巴地死了。解剖死鼠,发现有大量活的 S 型(而不是 R 型)细菌。他们推测,死细菌中的某一成分——转化源(transforming principle)将无致病力的细菌转化成病原细菌。10 多年后,实验表明,DNA 就是转化源。死细菌 DNA 指导了这一可遗传的转化,从而导致了小鼠的死亡。这一重大发现轰动了整个生物界。因为在当时,生物学家认为,只有像蛋白质这样复杂的大分子才可以担当决定细胞生物学特性和遗传的重托。Avery 等人的工作另辟蹊径,在遗传学理论上树立了全新的观点——DNA 是遗传信息的载体。

我们再来看一看美国冷泉港卡耐基遗传学实验室科学家 Hershey 和他的学生 Chase 在 1952 年从事噬菌体侵染细菌的实验(图 1-2)。噬菌体专门寄生在细菌体内。它的头、尾外部都有由蛋白质组成的外壳,头内主要是 DNA。噬菌体侵染细菌的过程可以分为以下 5 个步骤:①噬菌体用尾部的末端(基片、尾丝)吸

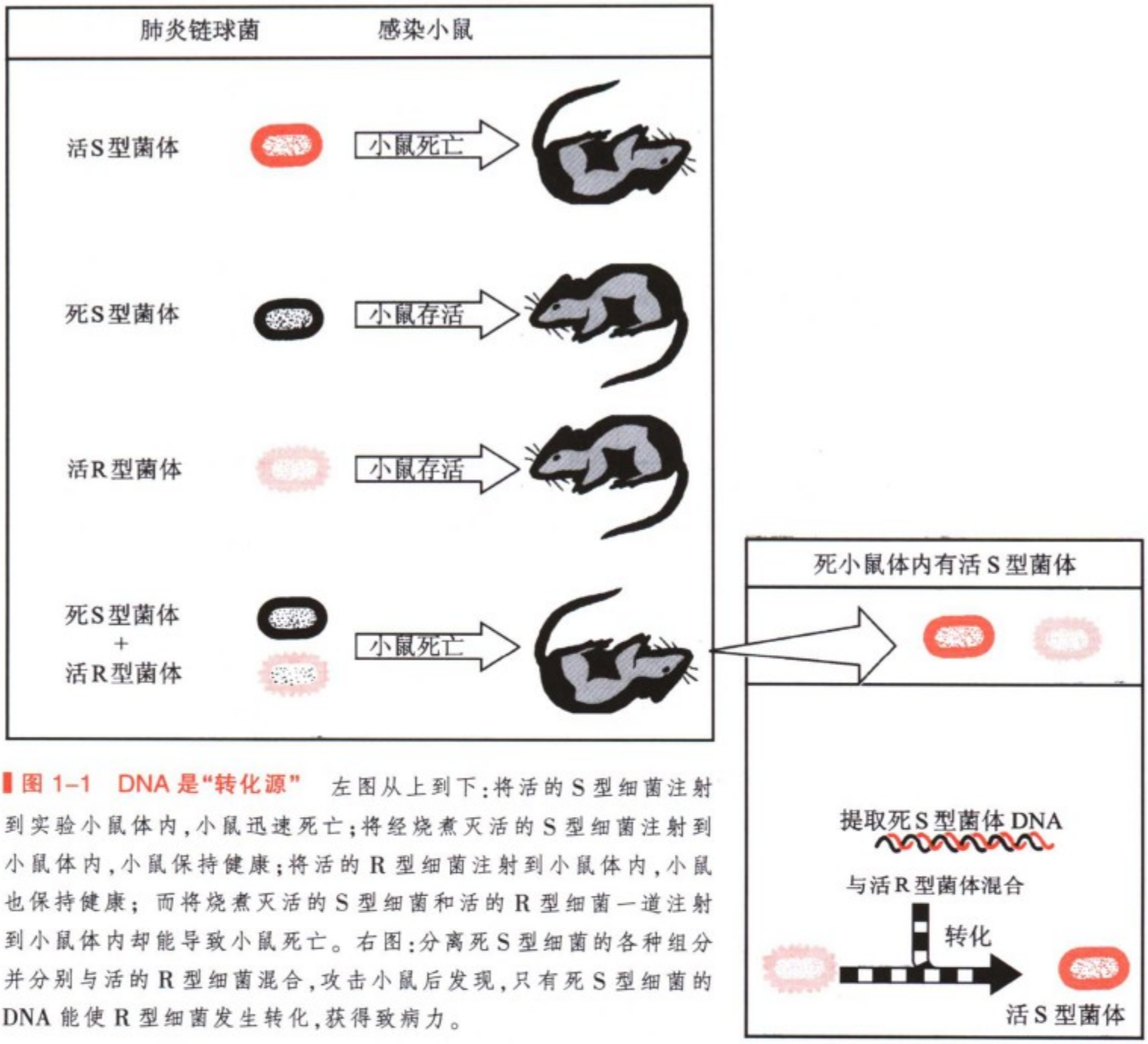



**图 1-1　DNA 是“转化源”**　左图从上到下：将活的 S 型细菌注射到实验小鼠体内，小鼠迅速死亡；将经烧煮灭活的 S 型细菌注射到小鼠体内，小鼠保持健康；将活的 R 型细菌注射到小鼠体内，小鼠也保持健康；而将烧煮灭活的 S 型细菌和活的 R 型细菌一道注射到小鼠体内却能导致小鼠死亡。右图：分离死 S 型细菌的各种组分并分别与活的 R 型细菌混合，攻击小鼠后发现，只有死 S 型细菌的 DNA 能使 R 型细菌发生转化，获得致病力。

附在细菌表面；②噬菌体通过尾轴把 DNA 全部注入细菌细胞内，噬菌体的蛋白质外壳则留在细胞外面；③噬菌体的 DNA 一旦进入细菌体内，它就能利用细菌的生命过程合成噬菌体自身的 DNA 和蛋白质；④新合成的 DNA 和蛋白质外壳，能组装成许许多多与亲代完全相同的子噬菌体；⑤子代噬菌体由于细菌的解体而被释放出来，再去侵染其他细菌。在整个过程中，DNA 起了关键作用。

当细菌培养基中分别带有 $^{35}S$ 或 $^{32}P$ 标记的氨基酸或核苷酸，子代噬菌体中就相应含有 $^{35}S$ 标记的蛋白质或 $^{32}P$ 标记的核酸。分别用这些噬菌体感染没有放射性标记的细菌，经过 1~2 个噬菌体 DNA 复制周期后发现，子代噬菌体中几乎不含带 $^{35}S$ 标记的蛋白质，但含有 30%以上的 $^{32}P$ 标记，说明在噬菌体传代过程中发挥作用的可能是 DNA，而不是蛋白质。

那么，DNA 到底是什么样的呢？在 1944 年的研究报告中，Avery 这样写道：当溶液中乙醇的体积达到 9/10 时，有纤维状物质析出。如稍加搅动，这种物质便会像棉线绕在线轴上一样缠绕在硬棒上，溶液中的其他成分则以颗粒状沉淀

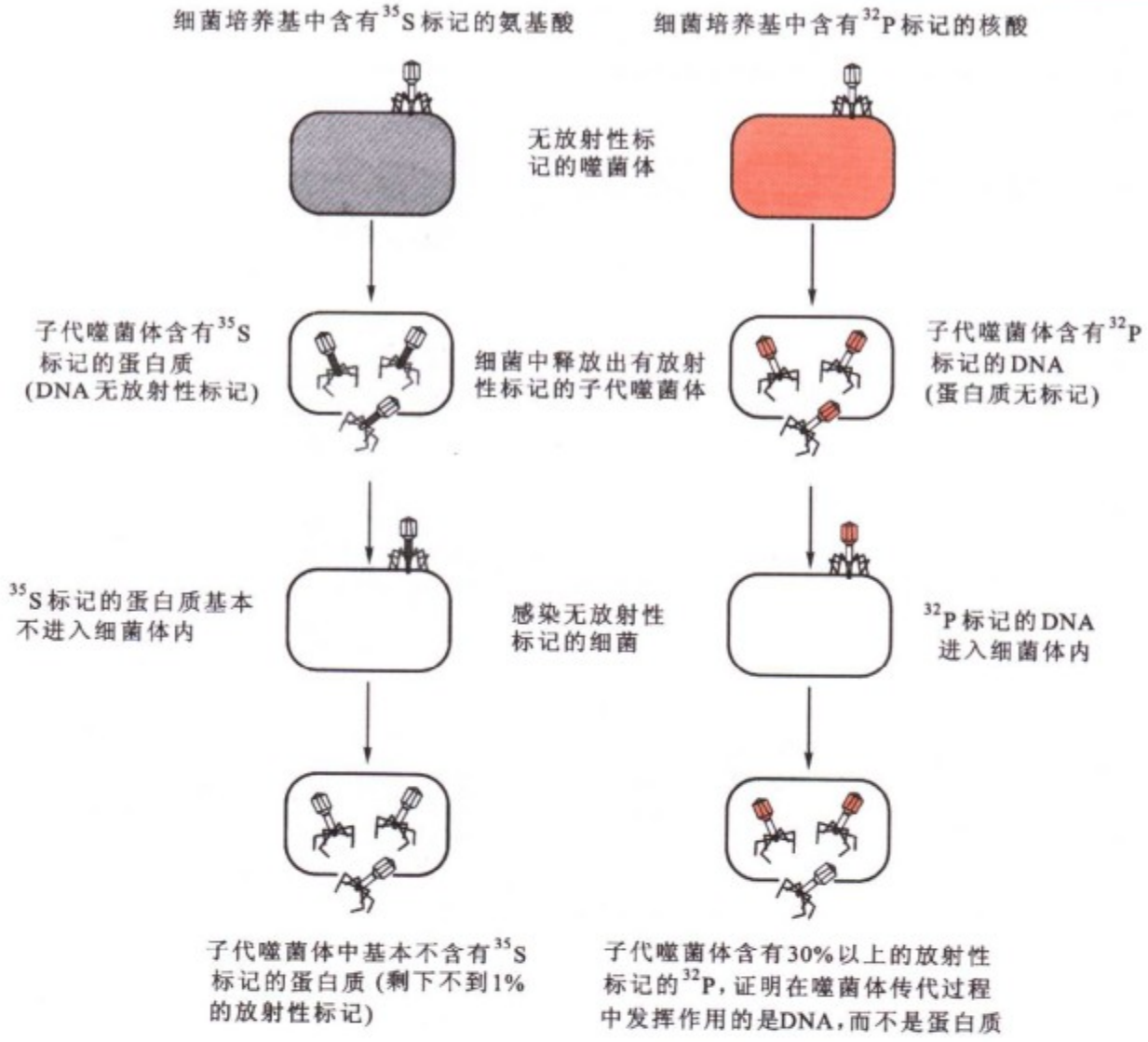



图 1-2 Hershey 和 Chase 证实噬菌体 DNA 侵染细菌的实验流程

留在下面。溶解纤维状物质并重复沉淀数次,可提高其纯度。这一物质具有很强的生物学活性,初步实验证实它很可能就是 DNA(谁能想到!)。对 DNA 分子的物理化学研究导致了现代生物学翻天覆地的革命,这更是 Avery 所没有想到的!

## 1.2 分子生物学简史

分子生物学是研究核酸、蛋白质等所有生物大分子的形态、结构特征及其重要性、规律性和相互关系的科学。它是人类从分子水平上真正揭开生物世界的奥秘,由被动地适应自然界转向主动地改造和重组自然界的基础学科。当人们意识到同一生物不同世代之间的连续性是由生物体自身所携带的遗传物质所决定的,科学家为揭示这些遗传密码所进行的努力就成为人类征服自然的一部分,而以生物大分子为研究对象的分子生物学就迅速成为现代生物学领域里最具活力的科学。

从 1847 年 Schleiden 和 Schwann 提出"细胞学说",证明动、植物都是由细胞组成的,到今天,虽然不过短短一百多年时间,我们对生物大分子——细胞的化学组成却有了深刻的认识。孟德尔的遗传学规律最先使人们对性状遗传产生

了理性认识，而 Morgan 的基因学说则进一步将“性状”与“基因”相偶联，成为现代遗传学的奠基石。随着核酸化学研究的进展，Watson 和 Crick 又提出了脱氧核糖核酸的双螺旋模型，为充分揭示遗传信息的传递规律铺平了道路。在蛋白质化学方面，继 Sumner 在 1936 年证实酶是蛋白质之后，Sanger 利用纸电泳及色谱技术于 1953 年首次阐明胰岛素的一级结构，开创了蛋白质序列分析的先河。而 Kendrew 和 Perutz 利用 X 线衍射技术解析了肌红蛋白(myoglobin)及血红蛋白(hemoglobin)的三维结构，论证了这些蛋白质在输送分子氧过程中的特殊作用，成为研究生物大分子空间构型的先驱。为了解分子生物学进程，我们不妨来看一看以部分诺贝尔生理医学奖和化学奖作为纽带的分子生物学发展简史。

1910 年，德国科学家 Kossel 因为蛋白质、细胞及细胞核化学的研究而获得诺贝尔生理医学奖，他首先分离出腺嘌呤、胸腺嘧啶和组氨酸。

1959 年，美籍西班牙裔科学家 Uchoa 发现了细菌的多核苷酸磷酸化酶，成功地合成了核糖核酸，研究并重建了将基因内的遗传信息通过 RNA 中间体翻译成蛋白质的过程。他和 Kornberg 分享了当年的诺贝尔生理医学奖，而后者的主要贡献在于实现了 DNA 分子在细菌细胞和试管内的复制。

1962 年，美国科学家 Watson 和英国科学家 Crick 因为在 1953 年提出 DNA 的反向平行双螺旋模型而与 Wilkins 共享诺贝尔生理医学奖，后者通过对 DNA 分子的 X 线衍射研究证实了 Watson 和 Crick 的 DNA 模型。

同年，英国科学家 Kendrew 和 Perutz 由于测定了肌红蛋白及血红蛋白的高级结构而荣获诺贝尔化学奖。

1965 年，法国科学家 Jacob 和 Monod 由于提出并证实了操纵子(operon)作为调节细菌细胞代谢的分子机制而与 Iwoff 分享了诺贝尔生理医学奖。除了著名的操纵子模型以外，Jacob 和 Monod 还首次提出存在一种与染色体脱氧核糖核酸序列相互补，能将编码在染色体 DNA 上的遗传信息带到蛋白质合成场所(细胞质)并翻译产生蛋白质的信使核糖核酸，即 mRNA 分子。他们的这一学说对分子生物学的发展起了极其重要的指导作用。

1968 年，美国科学家 Nirenberg 由于在破译 DNA 遗传密码方面的贡献，与 Holly 和 Khorana 等人分享了诺贝尔生理医学奖。Holly 的主要功绩在于阐明了酵母丙氨酸 tRNA 的核苷酸序列，并证实所有 tRNA 具有结构上的相似性。而 Khorana 第一个合成了核酸分子，并且人工复制了酵母基因。

1975 年，美国人 Temin、Dulbecco 和 Baltimore 由于发现在 RNA 肿瘤病毒中存在以 RNA 为模板，反转录生成 DNA 的反转录酶而共享诺贝尔生理医学奖。

1980 年，Sanger 因设计出一种测定 DNA 分子内核苷酸序列的方法，而与 Gilbert 和 Berg 分获诺贝尔化学奖。Berg 是研究 DNA 重组技术的元老，他最早(1972)获得了含有编码哺乳动物激素基因的工程菌株。Sanger 与 Gilbert 发明的 DNA 序列分析法至今仍被广泛使用，是分子生物学最重要的研究手段之一。此外，Sanger 还由于测定了牛胰岛素的一级结构而获得 1958 年的诺贝尔化学奖。

1983 年，美国遗传学家 McClintock 由于在 20 世纪 50 年代提出并发现了可移动的遗传因子(jumping gene，mobile element)而获得诺贝尔生理医学奖。

1984 年,德国人 Kohler、美国人 Milstein 和丹麦科学家 Jerne 由于发展了单克隆抗体技术,完善了极微量蛋白质的检测技术而分享了诺贝尔生理医学奖。

1989 年,美国科学家 Altman 和 Cech 由于发现某些 RNA 具有酶的功能(称为核酶)而共享诺贝尔化学奖。Bishop 和 Varmus 由于发现正常细胞同样带有原癌基因而分享当年的诺贝尔生理医学奖。

1993 年, 美国科学家 Roberts 和 Sharp 因其在断裂基因方面的研究成果而荣获诺贝尔生理医学奖。美国科学家 Mullis 由于发明 PCR 仪而与第一个设计基因定点突变的 Smith 共享诺贝尔化学奖。

1994 年, 美国科学家 Gilman 和 Rodbell 由于发现了 G 蛋白在细胞内信息转导中的作用而分享诺贝尔生理医学奖。

1995 年,美国人 Lewis、德国人 Nusslein-Volhard 和美国人 Wieschaus 由于在 20 世纪 40 至 70 年代先后独立鉴定了控制果蝇体节发育基因而分享诺贝尔生理医学奖。

1996 年,澳大利亚科学家 Doherty 和瑞士人 Zinkernagel 由于阐明了 T 淋巴细胞的免疫机制而分享了当年的诺贝尔生理医学奖。他们发现,白细胞只有同时识别入侵病原物和与之相伴的主要组织不相容抗原,才能准确识别受病原侵害的细胞并将其清除掉。

1997 年,美国科学家 Prusiner 由于发现朊病毒(prions)是阿尔茨海默病(早老性痴呆症)等疾病的病原并能直接在宿主细胞中繁殖传播而获得诺贝尔生理医学奖。

1999 年,美国科学家 Blobel 由于阐述了蛋白质在细胞间的运转机制,明确了信号肽及信号识别复合物(signal-recognition particles,SRPS)在蛋白质跨膜运转过程中的主导作用而获得诺贝尔生理医学奖。

2001 年,美国科学家 Hartwell、英国科学家 Hunt 和 Nurse 因为对细胞周期调控因子的研究而分享诺贝尔生理医学奖。

2006 年, 美国科学家 Kornberg 由于在揭示真核细胞转录机制方面的杰出贡献获得诺贝尔化学奖。美国科学家 Fire 和 Mello 由于在揭示控制遗传信息流动的基本机制——RNA 干扰方面的杰出贡献而获得诺贝尔生理医学奖。

此外,Avery 等人(1944)关于强致病性光滑型(S 型)肺炎链球菌 DNA 导致无毒株粗糙型(R 型)细菌发生遗传转化的实验,Meselson 和 Stahl(1958)关于 DNA 半保留复制的实验,Crick 于 1954 年所提出的遗传信息传递规律 (即中心法则,图1-3),Yanofsky 和 Brener(1961)关于遗传密码三联子的设想都对分子生物学的发展起了重大作用,将永载史册。

我国生物科学家吴宪于 20 世纪 20 年代初回国后,在协和医科大学生化系与汪猷、张昌颖等人一道完成了蛋白质变性理论、血液生化检测和免疫化学等一系列有重大影响的研究,成为我国生物化学界的先驱。20 世纪 60 年代、70 年代和 80 年代, 我国科学家相继实现了人工全合成有生物学活性的结晶牛胰岛素,解出了三方二锌猪胰岛素的晶体结构,采用有机合成与酶促相结合的方法完成了酵母丙氨酸转移核糖核酸的人工全合成,在酶学研究、蛋白质结构及生

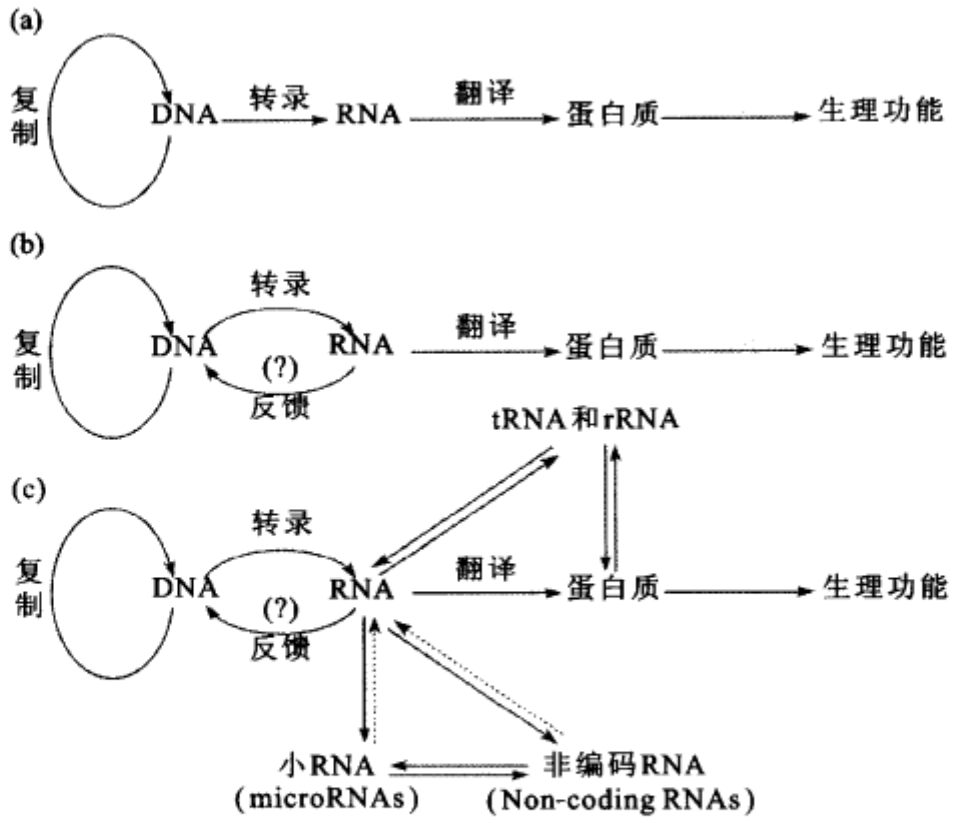



**图 1-3 遗传信息传递的中心法则** (a)Crick 在 1954 年首次提出的"中心法则"草图;(b)20 世纪 70 至 80 年代广为流传的"中心法则"示意图;(c)进入 21 世纪后,修正的"中心法则"示意图。

物膜结构与功能等方面都有世所瞩目的建树。

## 1.3 分子生物学的主要研究内容

现代生物学研究发现,所有生物体中的有机大分子都是以碳原子为核心,并以共价键的形式与氢、氧、氮及磷等以不同方式构成的。不仅如此,一切生物体中的各类有机大分子都是由完全相同的单体,如蛋白质分子中的 20 种氨基酸、DNA 及 RNA 中的 8 种碱基所组合而成的,由此产生了分子生物学的 3 条基本原理:

① 构成生物体各类有机大分子的单体在不同生物中都是相同的。

② 生物体内一切有机大分子的构成都遵循共同的规则。

③ 某一特定生物体所拥有的核酸及蛋白质分子决定了它的属性。

图 1-4 比较了生物体内各种大分子、亚细胞结构及原核、真核细胞的大小,了解并熟记它们的相对体积对于我们深入领会分子生物学所研究的生命过程有着不容忽视的作用。如在 B-DNA 中每个碱基对中心距离为 0.34nm,DNA 双螺旋的直径为 1.9 nm,而一个相对分子质量为 $5.0\times10^4$ 的球状蛋白的直径则为 5.0 nm,大约仅相当于 15~20 bp。核小体(nucleosome)由于外表面缠绕了 146 个碱基对,其相对分子质量则达到 $3.0\times10^5$,直径为 11nm。负责蛋白质合成的核糖体,其相对分子质量高达数百万,直径为 20 nm 以上。在 1 个体积为 1~2 $\mu m^3$ 的细菌细胞内,多聚核糖体确实可以算是一个不小的工厂了。

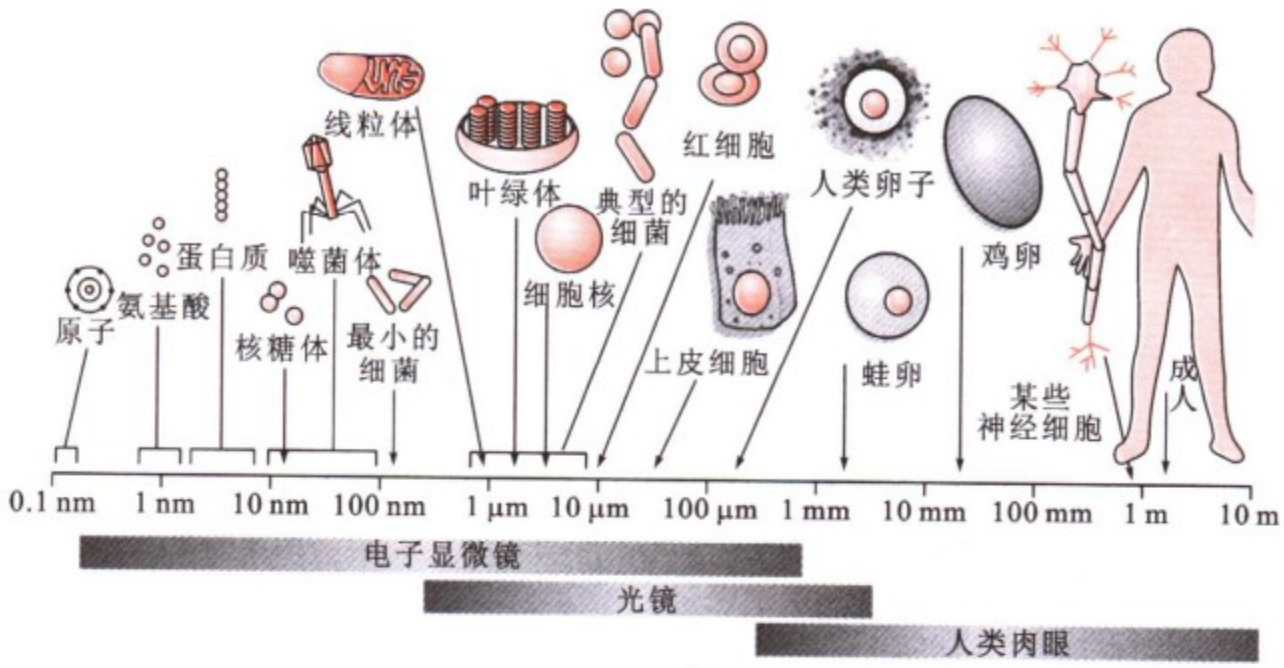



**图 1-4 生物世界的长度和体积比较** (a)生物界各种特征性生物大分子、各种亚细胞结构与细胞及成人的长度比较;(b)核酸、蛋白质、RNA 聚合酶核小体、核糖体和人染色体等主要生物大分子的体积比较。

从表面看,分子生物学涉猎范围极为广阔,研究内容也似乎包罗万象。事实上,它所研究的不外乎以下四个方面。

### 1.3.1 DNA 重组技术

DNA 重组技术又称基因工程,是 20 世纪 70 年代初兴起的技术科学,目的是将不同的 DNA 片段(如某个基因或基因的一部分)按照人们的设计定向连接起来,在特定的受体细胞中与载体同时复制并得到表达,产生影响受体细胞的新的遗传性状。严格地说,DNA 重组技术并不完全等于基因工程,因为后者还包括其他可能使生物细胞基因组结构得到改造的体系。DNA 重组技术是核酸化

学、蛋白质化学、酶工程及微生物学、遗传学、细胞学长期深入研究成果的结晶，而限制性内切酶、DNA 连接酶及其他工具酶的发现与应用则是这一技术得以建立的关键。

DNA 重组技术有着广阔的应用前景。首先，它可被用于大量生产某些在正常细胞代谢中产量很低的多肽，如激素、抗生素、酶类及抗体等，提高产量，降低成本，使许多有价值的多肽类物质得到广泛应用。由于发现了根癌农杆菌，发明了植物基因的轰击转化法，用转基因模式大规模改良农作物的抗病、抗逆、抗虫性，提高产量、改善品质或用传统农作物产生特种资源已经成为世界农业发展的潮流。

其次，DNA 重组技术可用于定向改造某些生物的基因组结构，使它们所具备的特殊经济价值或功能得以成百上千倍地提高。例如，有一种含有分解各种石油成分的重组 DNA 的超级细菌，能快速分解石油，可用来恢复被石油污染的海域或土壤。美国科学家应用该技术构建了“工程沙门菌”，在研制避孕疫苗方面取得了重要进展。首先去掉沙门菌致病基因部分，再引入来自精子的某些遗传信息，将改造后的细菌送入雌鼠体内，发现能产生排斥精细胞的抗体，使精子不能与卵子结合，从而达到避孕目的。美国陆军研究发展和工程中心还从织网蜘蛛中分离出合成蜘蛛丝的基因，并利用该基因在实验室里生产蜘蛛丝。将这一基因转移到细菌内，生产出一种可溶性丝蛋白，经浓缩后纺成一种强度超过钢的特殊纤维。研究人员希望对该基因进行修饰，以生产出高性能纤维，从而用于生产防弹背心、帽子、降落伞绳索和其他高强度的轻型装备。

最后，DNA 重组技术还被用来进行基础研究。如果说，分子生物学研究的核心是遗传信息的传递和控制，那么根据中心法则，我们要研究的就是从 DNA 到 RNA，再到蛋白质的全过程，也即基因的表达与调控。在这里，无论是对启动子的研究（包括调控元件或称顺式作用元件），还是对转录因子的克隆与分析，都离不开重组 DNA 技术的应用。

### 1.3.2 基因表达调控研究

因为蛋白质分子参与并控制了细胞的一切代谢活动，便决定蛋白质结构和合成时序的信息都由核酸（主要是脱氧核糖核酸）分子编码，表现为特定的核苷酸序列，所以基因表达实质上就是遗传信息的转录和翻译。在个体生长发育过程中，生物遗传信息的表达按一定的时序发生变化（时序调节），并随着内外环境的变化而不断加以修正（环境调控）。基因表达的调控主要发生在转录水平或翻译水平上。原核生物的基因组和染色体结构都比真核生物简单，转录和翻译在同一时间和空间内发生，基因表达的调控主要发生在转录水平。真核生物有细胞核结构，转录和翻译过程在时间和空间上都被分隔开，且在转录和翻译后都有复杂的信息加工过程，其基因表达的调控可以发生在各种不同的水平上。基因表达调控主要表现在信号传导研究、转录因子研究及 RNA 剪辑 3 个方面。信号传导是指外部信号通过细胞膜上的受体蛋白传到细胞内部，并激发诸如离子通透性、细胞形状或其他细胞功能的应答过程。当信号分子（配体）与相应的

受体作用后,可以激活受体分子的构型变化,使之形成专一性的离子通道,也可以激活受体分子的蛋白激酶或磷酸酯酶,还可以通过受体分子指导合成 cGMP、cAMP、肌醇三磷酸等第二信使分子。研究认为,信号传导之所以能引起细胞功能的改变,主要是由于信号最后活化了某些蛋白质分子,使之发生构型变化,从而直接作用于靶位点,打开或关闭某些基因。

转录因子是一群能与基因 5′端上游特定序列专一结合,从而保证目的基因以特定的强度在特定的时间与空间表达的蛋白质分子。在对植物的某些性状进行遗传分析时发现,某些基因的突变会影响其他基因的表达。例如,有 20 多个基因参与玉米花青素的生物合成,但其中的 *C1*、*r*、*pl* 或 *b* 基因发生突变后,该代谢途径中的结构酶基因全部被关闭。如果 *Antp*、*Flz* 或 *Ubx* 等基因发生突变,果蝇的体节发育就会受影响,身体中的一部分就可能变成相似于另一部分的结构,因此,它们是控制果蝇胚胎早期体节分化与发育的主要调节基因,它们所编码的蛋白质是调节与发育有关的结构基因表达的总开关。

真核基因在结构上的不连续性是近 10 年来生物学上的重大发现之一。当基因转录成 pre-mRNA 后,除了在 5′端加帽及 3′端加多聚 A(poly A)之外,还要切去隔开各个相邻编码区的内含子,使外显子(编码区)相连后成为成熟 mRNA。研究发现,许多基因中的内含子并不是一次全部切去,而是在不同的细胞或不同的发育阶段选择性剪切其中部分内含子,生成不同的 mRNA 及蛋白质分子。如降钙素基因、肌原蛋白基因和参与果蝇体细胞分化的 *dsx* 基因等,都采用选择性剪切方式从而生成不同功能的蛋白质。由于 RNA 的选择性剪切不牵涉遗传信息的永久性改变,所以是真核基因表达调控中一种比较灵活的方式。

### 1.3.3 生物大分子的结构功能研究——结构分子生物学

一个生物大分子,无论是核酸、蛋白质或多糖,在发挥生物学功能时,必须具备两个前提。首先,它拥有特定的空间结构(三维结构);其次,在它发挥生物学功能的过程中必定存在着结构和构象的变化。结构分子生物学就是研究生物大分子特定的空间结构及结构的运动变化与其生物学功能关系的科学。它包括结构的测定、结构运动变化规律的探索及结构与功能相互关系的建立 3 个主要研究方向。目前,最常见的研究三维结构及其运动规律的手段是 X 线衍射的晶体学(又称蛋白质晶体学),其次是用二维或多维核磁共振研究液相结构,也有人用电镜三维重组、电子衍射、中子衍射和各种频谱学方法研究生物高分子的空间结构。

### 1.3.4 基因组、功能基因组与生物信息学研究

2001 年 2 月,世界两大最著名的学术刊物,“Nature”和“Science”同时发表了人类基因组全序列,为确定基因对人类生长发育和疾病的预防治疗提供了一个前所未有的大舞台,生命科学界为之振奋。最新的数据表明,已有数十种原核生物及酵母、线虫、果蝇和高等植物拟南芥等多种真核生物基因组被基本破译,极大地丰富了人类的知识宝库,加快了人类认识自然和改造自然的步伐。虽然

完成某一生物的基因组计划就意味着该物种所有遗传密码已经为人类所掌握，但测定基因组序列只是了解基因的第一步，因为基因组计划不可能直接阐明基因的功能，更不能预测该基因所编码蛋白质的功能与活性，所以，并不能指导人们充分准确地利用这些基因的产物。于是，科学家又在基因组计划的基础上提出了“蛋白组计划”（又称“后基因组计划”或“功能基因组计划”），旨在快速、高效、大规模鉴定基因的产物和功能。

巨大的基因组信息给科学家带来了前所未遇的挑战。以人细胞中所带有的近 30 亿个碱基对为例，一个人即使每秒中读 10 个碱基，每天工作 24 小时，一年干 365 天，也得花 10 年时间才能把这些数据看一遍，更不用说数据分析了。依靠计算机快速高效运算并进行统计分类和结构功能预测的生物信息学就是在这样的背景下诞生的。没有生物信息学的知识，不借助于最先进的计算科学，人类就不可能最大限度地开发和运用基因组学所产生的庞大数据。

## 1.4 展望

从 20 世纪 50 年代初 Watson 和 Crick 提出 DNA 双螺旋模型至今，短短 40 多年间，生物学领域里的变化岂止“沧海桑田”所能形容。核苷酸序列测定技术的迅速进步，使人类基因组近 30 亿个碱基对全部被测定！20 年前，当人们第一次谈到这个巨大的项目时，不免带有“谈虎色变”的感觉。X 线衍射及其他高分子研究技术的相继问世，使建立生物大分子三维构象库的梦想成真。到 2006 年，已有近 30 000 套蛋白质构象入库。DNA 重组技术的应用，使得基因克隆分析日益成为全世界数以万计生物科学工作者手中的“常规武器”。近来，每年都有上千个新的基因序列被存入各类基因文库。根据 ISAAA（International Service for the Acquisition of Agri-biotech Applications）2007 年 1 月的统计资料，全球范围内已有 22 个国家正在种植转基因农作物，从 1996 年第一次进入大田种植以来的短短 11 年间，转基因农作物种植面积已经超过 $5.83\times10^6$ $km^2$，25 倍于英国全境。

20 世纪中期以来，生物学正在各个学科之间广泛渗透，相互促进，不断深入和发展，既从宏观和微观、最基本和最复杂等不同方向展开研究，也从分子水平、细胞水平、个体和群体等不同层次深入探索各种生物学现象，逐步揭开生命的奥秘。生物学革命也为数学、物理学、化学、信息、材料与工程科学提出了许多新概念、新问题和新思路，促使这些学科在理论和方法上得到发展提高。

生命世界的多样性和生命本质的一致性这个辩证的统一，已经为越来越多的人所接受。尽管生命过程在数以万计的不同生物中的表现形式可以是完全不同的，但生命活动的本质是高度一致的，如核酸与蛋白质一级结构的对应关系，在整个生命世界都是一致的。

除极少数生物体外，脱氧核糖酸是地球上千百万生灵所共有的遗传密码。如果没有这个统一性，人们就不可能把某一个基因从 A 生物转移到 B 生物体内，得到表达并发挥相同的功能。从表面上看，动物和植物是两个完全不同的群

体,它们以两种完全不同的方式摄取能量。动物靠的是氧化磷酸化,在食物的氧化过程中合成“生命通货”——腺苷三磷酸(ATP),而植物则通过光合作用,将光能转变成 ATP,以供生命活动之需。但是,动、植物细胞代谢活动的实质都是电子在一系列受体蛋白质之间传递,造成膜内外质子浓度差,以合成 ATP。生命活动的这种高度一致性,使分子生物学研究日益渗透到生物学的各个领域,产生了全面的影响。

分子生物学、细胞生物学和神经生物学被认为是当代生物学研究的三大主题,分子生物学的全面渗透推动了细胞生物学和神经生物学的发展。分子生物学研究技术的发展,几乎完全改变了科学家对膜内外信号转导、离子通道的分子结构、功能特性及运转方式的认识。在对突触部位神经递质的合成、维持、释放及其作用的分子机制的研究中,最近 10 年所取得的进展远远超过了以往几十年的总和。

遗传学是分子生物学发展以来受影响最大的学科。孟德尔著名的皱皮豌豆和圆粒豌豆子代分离实验以及由此得到的遗传规律,纷纷在近 20 年内得到分子水平上的解释。越来越多的遗传学原理正在被分子水平的实验所证实或否定,许多遗传病已经得到控制或可以治愈,许多经典遗传学无法解决的问题和无法破译的奥秘,也相继被攻克,分子遗传学已成为人类了解、阐明和改造自然界的重要武器。

分类和进化研究是生物学中最古老的领域,它们同样由于分子生物学的渗透而获得了新生。过去研究分类和进化,主要依靠生物体的形态,并辅以生理特征,来探讨生物间亲缘关系的远近。现在,反映不同生命活动中更为本质的核酸、蛋白质序列间的比较,已被大量用于分类和进化的研究。由于核酸技术的进步,科学家已经可能从已灭绝的化石里提取极为微量的 DNA 分子,并进行深入的研究,以此确立这些生物在进化树中的地位。

分子生物学还对发育生物学研究产生了巨大的影响。人们早就知道,个体生长发育所需的全部信息都是储存在 DNA 序列中的,如果受精卵中的遗传信息不能按照一定的时空顺序表达,个体发育规律就会被打乱,高度有序的生物世界就不复存在。大量分子水平的实验证明,非编码 RNA (包括 miRNA 和 siRNA)在动植物个体发育过程中发挥了举足轻重的作用。专家估计,这个领域的研究将为发育生物学带来一场革命。

分子生物学的发展揭示了生命本质的高度有序性和一致性,是人类在认识论上的重大飞跃。生命活动的一致性,决定了二十一世纪的生物学将是真正的系统生物学(systems biology),是生物学范围内所有学科在分子水平上的统一。以基因组学、转录组学、蛋白质组学以及代谢组学等不同层次“组学”的最新成果为基础的系统生物学,是研究一个生物系统中所有组成成分(基因、mRNA、蛋白质等)的变化规律以及在特定遗传或环境条件下相互关系的学科。系统生物学研究通过整合各组分的信息,以图画或数学方式建立能描述系统结构和行为的模型。由于分子生物学、生物化学及生物物理学的影响,大量物理、化学工作者进入生物学领域,既有力地推动了生命科学的发展,也极大地促进了这两个

学科的发展。因此，分子生物学不仅是目前自然科学中进展最迅速、最具活力和生气的领域，也是新世纪的带头学科。愿有志于生命科学研究的莘莘学子及广大青年朋友们，早日掌握分子生物学这把金钥匙，为祖国繁荣富强，为人类的文明和进步作出新贡献！

## 思考题

1. 简述孟德尔、摩尔根和沃森等人对分子生物学发展的主要贡献。
2. 写出 DNA 和 RNA 的英文全称。
3. 试述"有其父必有其子"的生物学本质。
4. 早期主要由哪些实验证实 DNA 是遗传物质？写出这些实验的主要步骤。
5. 请定义 DNA 重组技术和基因工程技术。
6. 说出分子生物学的主要研究内容。
7. 你认为二十一世纪初分子生物学将在哪些领域取得进展？

## 主要参考文献


1. Avery T O, MacLeod C M, McCarty M. Studies of the chemical nature of the substance inducing transformation of pneumococcal types. J. Exp. Med. 1944, 79: 137–158.
2. Costa F F. Non-coding RNAs: New players in eukaryotic biology . Gene, 2005, 357: 83–94.
3. Crick F H, Barnett L, Brenner S et al. General nature of the genetic code for proteins. Nature, 1961, 192:1227–1232.
4. Hershey A D, Chase M. Independent functions of viral protein and nucleic acid in growth of bacteriophages. J. Gen. Physiol. 1952, 36:39–56.
5. Meselson M, Stahl F W. The replication of DNA in Escherichia coli. Proc. Nat. Acad. Sci. U.S.A. 1958, 44:671–682.
6. Watson J D, Crick F H. A structure for deoxyribose nucleic acid. Nature, 1953, 171: 737–738.
7. Watson J D, Crick F H. Genetic implications of the structure of deoxyribonucleic acid. Nature, 1953, 171:964–967.
8. Wilkins M F H, Stokes A R, Wilson H R. Molecular structure of deoxypentose nucleic acids. Nature, 1953, 171:738–740.

# 2 染色体与 DNA


- 2.1 染色体
- 2.2 DNA 的结构
- 2.3 DNA 的复制
- 2.4 原核生物和真核生物 DNA 复制的特点
- 2.5 DNA 的修复
- 2.6 DNA 的转座


很久以前，人们就知道所有生命体都具备代代相传的本领。然而，生物遗传的物质基础是什么？或者说，什么因素决定了生物遗传呢？现代遗传学，尤其是分子生物学的研究证实，脱氧核糖核酸，简称 DNA，控制了生物的性状遗传（也有以核糖核酸——RNA 作为遗传物质的，如部分病毒和类病毒）。无论 DNA 或RNA，都是由许许多多个核苷酸连接而成的生物大分子，而每个核苷酸又由磷酸、核糖和碱基 3 部分组成（图 2-1）。组成 DNA 分子的碱基只有 4 种，即腺嘌呤(A)，鸟嘌呤(G)胸腺嘧啶(T)和胞嘧啶(C)。图 2-2 是存在于 DNA 和 RNA 分子中的 5 种含氮碱基的结构式。

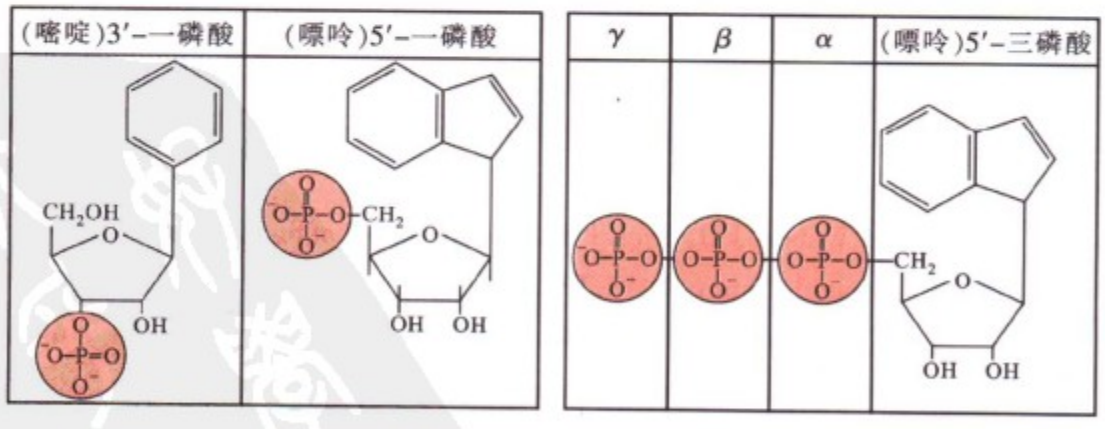



**图 2-1 核苷酸的组成与结构** 核苷的 5′位和 3′位都可能与磷酸基团相连。

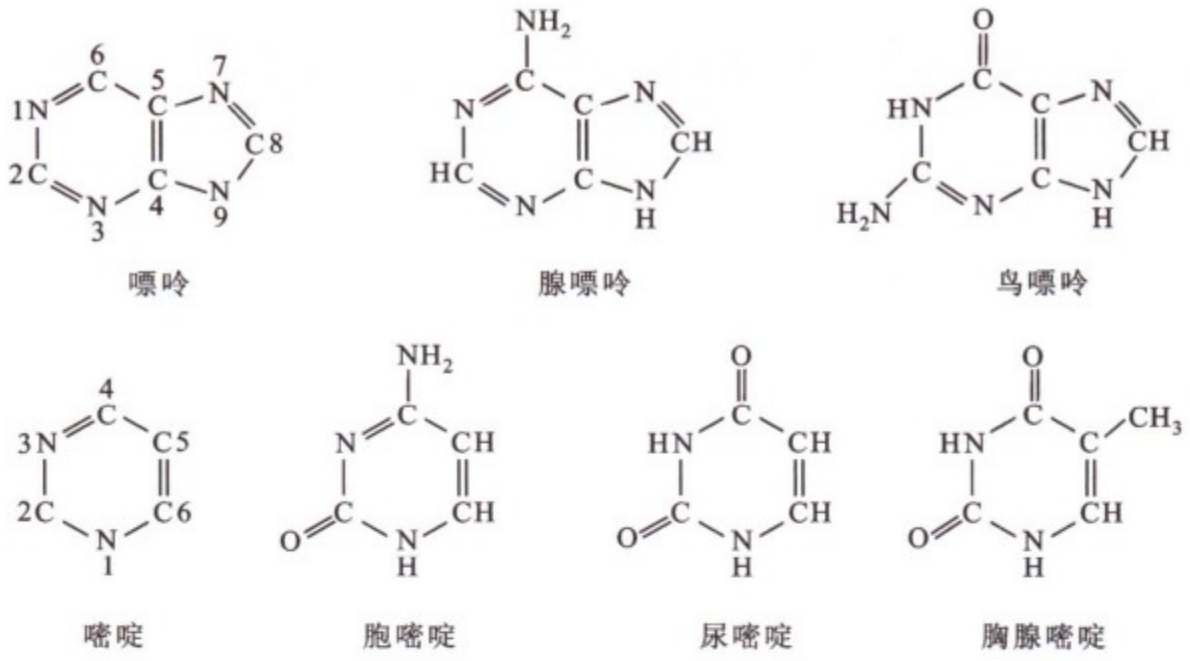



图 2-2 组成 DNA 和 RNA 分子的五种含氮碱基的结构式

# 2.1 染色体

## 2.1.1 染色体概述

由于亲代能够将自己的遗传物质 DNA 以染色体(chromosome)的形式传给子代,保持了物种的稳定性和连续性。因此,人们普遍认为染色体在遗传上起着主要作用。染色体包括 DNA 和蛋白质两大部分。同一物种内每条染色体所带 DNA 的量是一定的,但不同染色体或不同物种之间变化很大,从上百万到几亿个核苷酸不等。人 X 染色体就带有 1.28 亿个核苷酸对,而 Y 染色体只带有 0.19 亿个核苷酸对(表 2-1)。此外,组成染色体的蛋白质(组蛋白和非组蛋白)种类和含量也是十分稳定的。由于细胞内的 DNA 主要在染色体上,所以说遗传物质的主要载体是染色体。

表 2-1 人类基因组各条染色体中碱基对数量和推导的功能基因数量对照 *

| 染色体编号 | 长度/Mb | 估计的基因数 |
|---|---|---|
| 1 | 220 | 3 453 |
| 2 | 240 | 2 954 |
| 3 | 200 | 2 427 |
| 4 | 186 | 1 861 |
| 5 | 182 | 2 136 |
| 6 | 172 | 2 257 |
| 7 | 146 | 1 831 |
| 8 | 146 | 1 560 |
| 9 | 113 | 1 537 |
| 10 | 130 | 1 653 |

续表

| 染色体编号 | 长度/Mb | 估计的基因数 |
|---|---|---|
| 11 | 132 | 2 185 |
| 12 | 134 | 1 861 |
| 13 | 99 | 1 032 |
| 14 | 87 | 1 283 |
| 15 | 80 | 1 198 |
| 16 | 75 | 1 421 |
| 17 | 78 | 1 545 |
| 18 | 79 | 826 |
| 19 | 58 | 1 675 |
| 20 | 61 | 986 |
| 21 | 33 | 449 |
| 22 | 36 | 835 |
| X | 128 | 1 465 |
| Y | 19 | 210 |
| 总长 | 2 907 | 39 114 |

* 引自 Science. 2001, 291: 1304 ~ 1351

19 世纪中叶，细胞生物学家首先在光学显微镜下看到了细胞这个生命的基本单位(表 2-2)，他们不但记录了真核细胞分裂的全过程，发现了细胞质与细胞核之间的界限，还在细胞分裂时观察到存在于细胞核中的棒状可染色结构并将其命名为染色体。一般说来，只有在细胞有丝分裂过程中，人们才能在光学显微镜下观察到这些结构。

表 2-2 原核和真核细胞特征分析

| 特 征 | 原 核 | 真 核 |
|---|---|---|
| 体积 | 小(1 ~ 10 μm) | 较大(5 ~ 100 μm) |
| 基因组 | DNA 与非组蛋白相结合，基因组位于类核中 | DNA 与组蛋白及非组蛋白结合形成染色体 |
| 细胞分裂形式 | 裂殖或出芽 | 有丝分裂和减数分裂 |
| 外膜包被的细胞器 | 无 | 线粒体、叶绿体、内质网、高尔基体等 |
| 能量代谢 | 无线粒体，氧化还原酶系统位于细胞质膜上 | 氧化还原酶系统位于线粒体中，能量代谢途径之间差异较小 |
| 细胞骨架 | 无 | 由微管、中间纤维和肌动纤维等形成复杂的细胞骨架 |
| 细胞间的运动 | 无 | 原生质体流动或胞饮 |

染色体位于真核细胞核的核仁内，是极细微的线性构造，因为它控制了生命遗传，所以人们又称之为“生命之线”。当细胞分裂时，每一条染色体都复制生成一条与母链完全一样的子链，形成同源染色体对。一般说来，染色体只有在细

胞有丝分裂过程中,才可在光学显微镜下观察到(图 2-3)。而在细胞生活周期中占较长时间的分裂间期，染色体以较细且松散的染色质形式存在于胞核中。非分裂期的细胞核经低渗处理、溶胀破裂后释放出染色质,在电子显微镜下呈纤维串珠状的长丝,包括有 DNA 双螺旋的长臂及部分蛋白质。

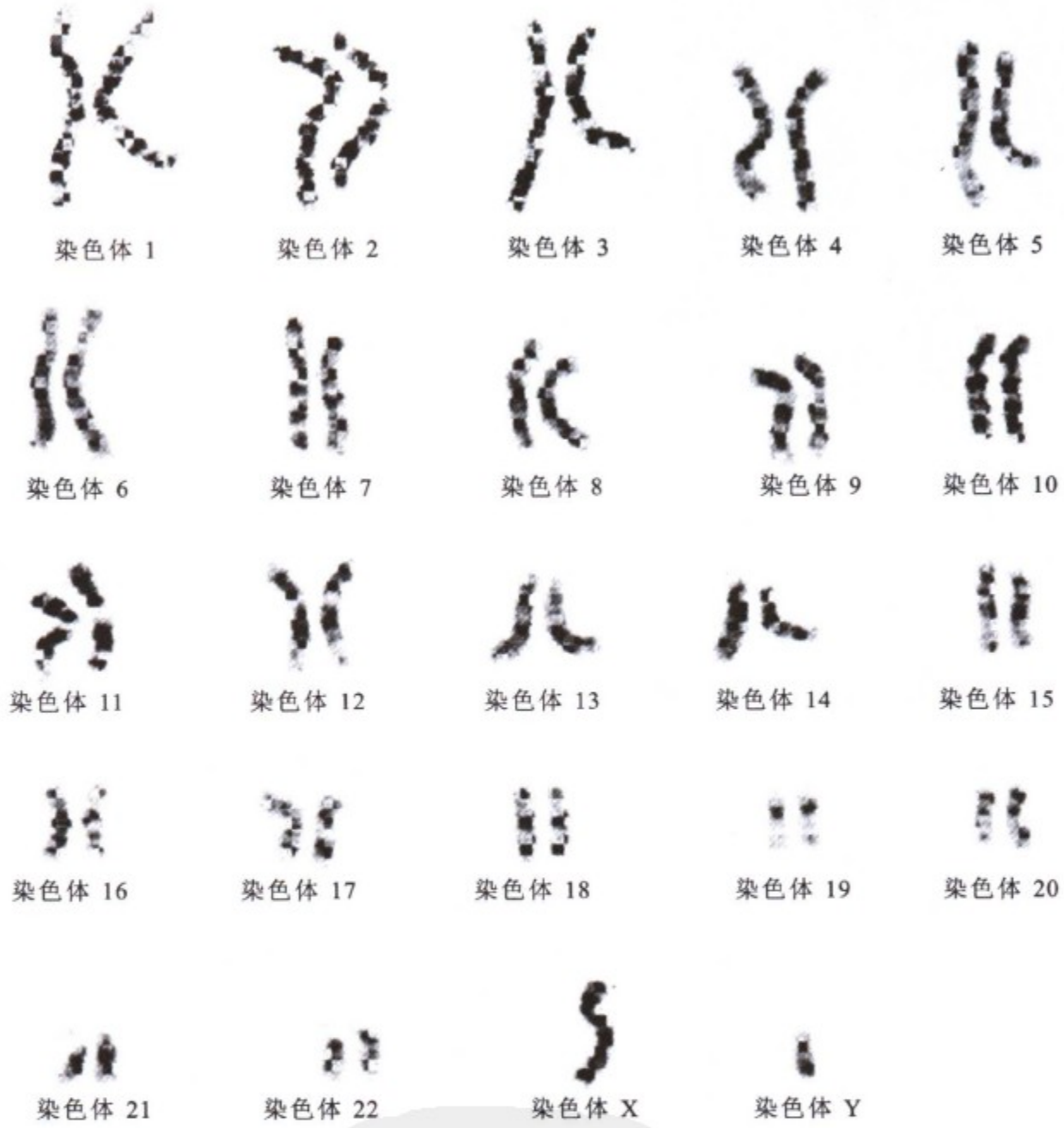



**图 2-3 人类 22 对染色体及性染色体显微结构图**

真核与原核细胞染色体除了外观结构不同之外,它们在细胞内的存在方式也不同。细菌染色体外裹着稀疏的蛋白质,其中的一部分与 DNA 的折叠有关,另一些则参与 DNA 复制、重组及转录过程。真核细胞的染色体中,DNA 与蛋白质完全融合在一起,其蛋白质与相应 DNA 的质量比约为 2:1。这些蛋白质,包括组蛋白和非组蛋白,在染色体的结构中起着重要作用。这样,DNA、组蛋白和非组蛋白及部分 RNA (主要是尚未完成转录而仍与模板 DNA 相连接的那些 RNA,其含量不到 DNA 的 10%)组成了染色体。

因为原核生物没有真正的细胞核,DNA 一般位于一个类似“核”的结构——称为类核体上。细菌 DNA 是一条相对分子质量在 $10^9$ 左右的共价、闭

合双链分子,通常也称为染色体(图 2-4)。虽然快速生长期内的大肠杆菌可以有几条染色体,但一般情况下只含有一条染色体。因此,大肠杆菌和其他原核细胞都是单倍的。

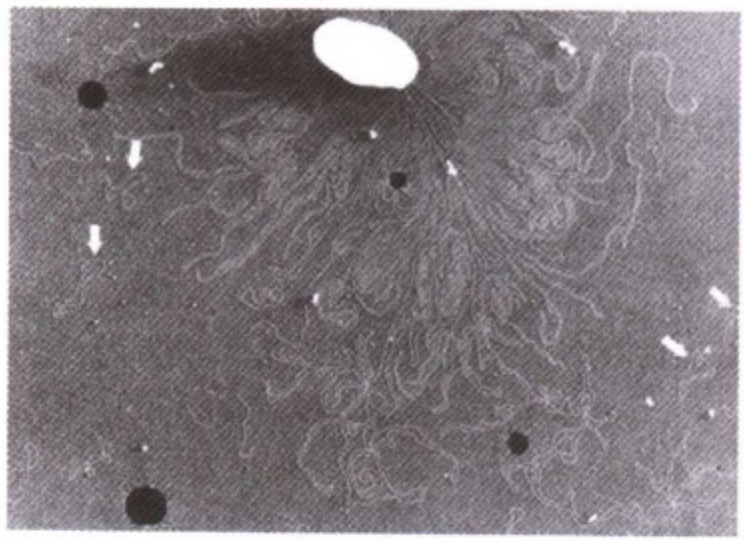


**图 2-4 大肠杆菌细胞中基因组 DNA 的电镜显微照片** 箭头所示处为球状质粒 DNA,各种白斑或黑斑为样品准备中产生的假象

原核生物 DNA 的主要特征:原核生物中一般只有一条染色体且大都带有单拷贝基因,只有很少数基因(如 rRNA 基因)是以多拷贝形式存在;整个染色体 DNA 几乎全部由功能基因与调控序列所组成;几乎每个基因序列都与它所编码的蛋白质序列呈线性对应状态。

### 2.1.2 真核细胞染色体的组成

真核细胞都有明显的核结构,除了性细胞以外,真核细胞的染色体都是二倍体,而性细胞即生殖细胞的染色体数目是体细胞的一半,故称为单倍体。在二倍体阶段,每个基因都有两个拷贝,其中的一个拷贝会通过配子(gamete),即性细胞(精子或卵子)从亲本传到子代。精、卵细胞结合形成合子(zygote),也称受精卵,它包含了从父、母本来的各一个拷贝的所有基因,从而创造了一个新生的二倍体。在这个新的个体内,基因的一个拷贝来自父本,另一个拷贝来自母本。

作为遗传物质,染色体具有如下特征:①分子结构相对稳定;②能够自我复制,使亲子代之间保持连续性;③能够指导蛋白质的合成,从而控制整个生命过程;④能够产生可遗传的变异。

表 2-3 给出了主要动、植物的染色体数目,表 2-4 是几种噬菌体的基因组 DNA 和病毒颗粒大小比较。

由于真核细胞 DNA 相对分子质量一般大大超过原核生物,其染色体也常常为大量蛋白质及核膜所包围,DNA 的转录和翻译是在不同的空间和时间上进行的,所以其基因表达的调控不仅与 DNA 的序列有关,而且也与染色体的结构有关。下面将详细介绍真核细胞染色体的组成。

1. 蛋白质

染色体上的蛋白质主要包括组蛋白和非组蛋白。组蛋白是染色体的结构蛋白,它与 DNA 组成核小体。通常可以用 2 mol/L NaCl 或 0.25 mol/L 的 $HCl/H_2SO_4$ 处理染色质使组蛋白与 DNA 分开,然后再用离子交换柱色谱分离。根据其凝胶

表 2-3 部分动、植物细胞内染色体数

| 物 种 | 单倍体染色体数 | 物 种 | 单倍体染色体数 |
|---|---|---|---|
| 人 | 23 | 线虫 | 11/12(M/F) |
| 猿猴 | 21 | 酵母 | 8~16 |
| 牛 | 30 | 绿藻 | 10 左右 |
| 狗 | 39 | 洋葱 | 8 |
| 猪 | 19 | 大麦 | 7 |
| 马 | 32 | 水稻 | 12 |
| 家鼠 | 20 | 玉米 | 10 |
| 大鼠 | 21 | 龙头花 | 8 |
| 荷兰猪 | 32 | 番茄 | 12 |
| 兔 | 22 | 烟草 | 24 |
| 鸡 | 39 | 菜豆 | 11 |
| 鳄鱼 | 16 | 松树 | 12 |
| 蛙 | 13 | 豌豆 | 7 |
| 鲤鱼 | 52 | 马铃薯 | 24 |
| 蚕 | 28 | 家蝇 | 6 |
| 淡水鳃 | 16 | 果蝇 | 4 |
| 狐 | 17 | 细菌 | 1 |
| 猫 | 19 | 红三叶草 | 7 |

注:有些生物染色体数并不固定,有些生物雌、雄个体间染色体数不同。

表 2-4 部分噬菌体基因组 DNA 及病毒颗粒大小比较

| 病 毒 | DNA 的长度/bp | DNA 的长度/nm | 病毒颗粒的长度/nm |
|---|---|---|---|
| φX174 | 5 386 | 1 939 | 25 |
| T 7 | 39 936 | 14 377 | 78 |
| λ(lambda) | 48 502 | 17 460 | 190 |
| T 4 | 168 889 | 60 800 | 210 |

电泳性质可以把组蛋白分为 $H_1$、$H_2A$、$H_2B$、$H_3$ 及 $H_4$。这些组蛋白都含有大量的赖氨酸和精氨酸,其中 $H_3$、$H_4$ 富含精氨酸,$H_1$ 富含赖氨酸;$H_2A$、$H_2B$ 介于两者之间。它们的理化特性见表 2-5。

表 2-5 真核细胞染色体上的组蛋白成分分析

| 种类 | 相对分子质量 | 氨基酸数目 | 分离难易度 | 保守性 | 染色质中比例 | 染色质中位置 |
|---|---|---|---|---|---|---|
| $H_1$ | 21 000 | 223 | 易 | 不保守 | 0.5 | 接头 |
| $H_2A$ | 14 500 | 129 | 较难 | 较保守 | 1 | 核心 |
| $H_2B$ | 13 800 | 125 | 较难 | 较保守 | 1 | 核心 |
| $H_3$ | 15 300 | 135 | 最难 | 最保守 | 1 | 核心 |
| $H_4$ | 11 300 | 102 | 最难 | 最保守 | 1 | 核心 |

组蛋白具有如下特性：

① 进化上的极端保守性　不同种生物组蛋白的氨基酸组成是十分相似的，特别是 $H_3$、$H_4$。牛、猪、大鼠的 $H_4$ 氨基酸序列完全相同。牛的 $H_4$ 序列与豌豆序列相比只有两个氨基酸的差异（豌豆 $H_4$ 中的异亮氨酸 $^{60}$→缬氨酸 $^{60}$，精氨酸 $^{77}$→赖氨酸 $^{77}$）。$H_3$ 的保守性也很强，鲤鱼与小牛胸腺的 $H_3$ 只差一个氨基酸，小牛胸腺与豆苗 $H_3$ 只差 4 个氨基酸。$H_2A$、$H_2B$ 的变化相对大些，$H_1$ 的变化则更大。$H_3$ 及 $H_4$ 在氨基酸组成上的极端保守性表明，它们可能对稳定真核生物的染色体结构起重要作用。

② 无组织特异性　到目前为止，仅发现鸟类、鱼类及两栖类红细胞染色体不含 $H_1$ 而带有 $H_5$，精细胞染色体的组蛋白是鱼精蛋白这两个例外。

③ 肽链上氨基酸分布的不对称性　碱性氨基酸集中分布在 N 端的半条链上。例如，$H_4$N 端的半条链上净电荷为+16，C 端只有+3，大部分疏水基团都分布在 C 端。这种不对称的分布可能与它们的功能和相互作用有关。碱性的半条链易与 DNA 的负电荷区结合，而另外半条链与其他组蛋白、非组蛋白结合。表 2-6 比较了不同组蛋白中赖氨酸和精氨酸的含量。

表 2-6　不同组蛋白分子中所含的碱性氨基酸比较（占氨基酸总量的%）

| 碱性氨基酸 | $H_1$ | $H_2A$ | $H_2B$ | $H_3$ | $H_4$ |
|---|---|---|---|---|---|
| 赖氨酸 | 29.5 | 10.9 | 16.0 | 9.6 | 10.8 |
| 精氨酸 | 1.3 | 9.3 | 6.4 | 13.3 | 13.7 |

④ 组蛋白的修饰作用　包括甲基化、乙基化、磷酸化及 ADP 核糖基化等。在几种组蛋白中以 $H_3$、$H_4$ 的修饰作用比较普遍，$H_2A$、$H_2B$、$H_1$ 不很明显，其中 $H_2B$ 有乙酰化作用，$H_1$ 有磷酸化作用。修饰作用只发生在细胞周期的特定时间和组蛋白的特定位点上，如组培细胞 $H_1$ 的磷酸化都发生在 S 期，而其他细胞则发生在 $G_2$ 后期，$H_2B$ 的乙基化都发生在赖氨酸残基上。

⑤ 富含赖氨酸的组蛋白 $H_5$　从该组蛋白的氨基酸成分来看，除了富含赖氨酸(24%)以外，还富含丙氨酸(16%)、丝氨酸(13%)及精氨酸(11%)。从鸟类、两栖类及鱼类红细胞分离的 $H_5$ 都具有种特异性，从氨基酸序列分析看，它与 $H_1$ 并无明显的亲缘关系。有核的红细胞完全失去复制和转录的能力，其核很小，染色质高度浓缩，所以人们认为红细胞染色质的失活与 $H_5$ 的积累有关。但是，在成红细胞中也有相当数量的 $H_5$。有证据表明，$H_5$ 的磷酸化很可能在染色质失活过程中起重要作用。

染色体上除了存在大约与 DNA 等量的组蛋白以外，还存在大量的非组蛋白。非组蛋白大约占组蛋白总量的 60%~70%，它的种类很多，约 20~100 种，其中常见的有 15~20 种。它们的相对分子质量在 $1.5\times10^4$~$1.8\times10^5$。非组蛋白包括酶类，如 RNA 聚合酶及与细胞分裂有关的收缩蛋白、骨架蛋白、核孔复合物蛋白以及肌动蛋白、肌球蛋白、微管蛋白、原肌蛋白等，它们也可能是染色质的结构成分。

① 高速泳动蛋白（high mobility group protein，HMG 蛋白）　这是一类能用

低盐(0.35 mol/L NaCl)溶液抽提、能溶于2%的三氯乙酸、相对分子质量都在$3.0\times10^4$以下的非组蛋白。因其相对分子质量小、在凝胶电泳中迁移速度快而得名。其中HMG1及HMG2的相对分子质量分别为$2.75\times10^4$和$2.6\times10^4$,富含赖氨酸、精氨酸、谷氨酸和天冬氨酸。这类蛋白的特点是能与DNA结合,也能与$H_1$作用,但都很容易用低盐溶液抽提,说明它们与DNA的结合并不牢固。现在一般认为这类蛋白可能与DNA的超螺旋结构有关。

② DNA结合蛋白　用2 mol/L NaCl除去全部组蛋白和70%的非组蛋白后,还有一部分非组蛋白紧紧地与DNA结合在一起,只有用2 mol/L NaCl和5 mol/L尿素才能把这些蛋白解离。它们是一些相对分子质量较低的蛋白质,约占非组蛋白的20%,染色质的8%。它们可能是一些与DNA的复制或转录有关的酶或调节物质。

③ A24非组蛋白　有科学家用0.2 mol/L的硫酸从小鼠肝脏中分离到一种称为A24的非组蛋白。这种蛋白的溶解性质与组蛋白相似。氨基酸序列测定发现A24的C端与$H_2A$相同,但它有两个N端,一个N端的序列与$H_2A$相同,另一个N端与泛蛋白相同,是在$H_2A$的119位赖氨酸的$\varepsilon$-$NH_2$上附加的多肽。它与$H_2A$差不多大小,呈酸性,含有较多的谷氨酸和天冬氨酸。A24的总量大约是$H_2A$的1%,位于核小体内,功能不详。

2. 真核生物基因组DNA

真核细胞基因组的最大特点是它含有大量的重复序列，而且功能DNA序列大多被不编码蛋白质的非功能DNA所隔开。我们把一种生物单倍体基因组DNA的总量称为C值(C value)。在真核生物中,C值一般是随生物进化而增加的,高等生物的C值一般大于低等生物,这很容易被高等生物需要更多的基因来控制性状所解释。然而,进一步的实验发现,某些两栖动物的C值甚至比哺乳动物还大,而在两栖动物中C值的变化也很大,可相差100倍(图2-5),这就是著名的“C值反常现象(C-value paradox)”,也称C值谬误。它是指C值往往与种系进化的复杂程度不一致,某些低等生物却具有较大的C值。如果这些DNA都是编码蛋白质的功能基因,那么,很难想象在两个相近的物种中,它们的基因数目会相差100倍。由此推断,许多DNA序列可能不编码蛋白质,是没有生理功能的。根据对DNA的动力学研究。真核细胞DNA序列大致上可被分为3类:

(1) 不重复序列　在单倍体基因组里,这些序列一般只有一个或几个拷贝,它占DNA总量的40%~80%，如牛细胞中占55%，小鼠中占70%，果蝇中占79%。不重复序列长约750~2 000 bp,相当于一个结构基因的长度。实际上结构基因基本上属于不重复序列,如卵清蛋白、蚕的丝心蛋白、血红蛋白和珠蛋白等都是单拷贝基因。

单拷贝基因通过基因扩增仍可合成大量的蛋白质,如一个蚕丝心蛋白基因可作为模板合成$10^4$个丝心蛋白mRNA,每个mRNA可存活4d,共合成$10^5$个丝心蛋白,这样,在几天之内,一个单拷贝丝心蛋白基因就可以合成$10^9$个丝心蛋白分子。这种放大作用特别是第二步放大作用对单拷贝基因来说是十分

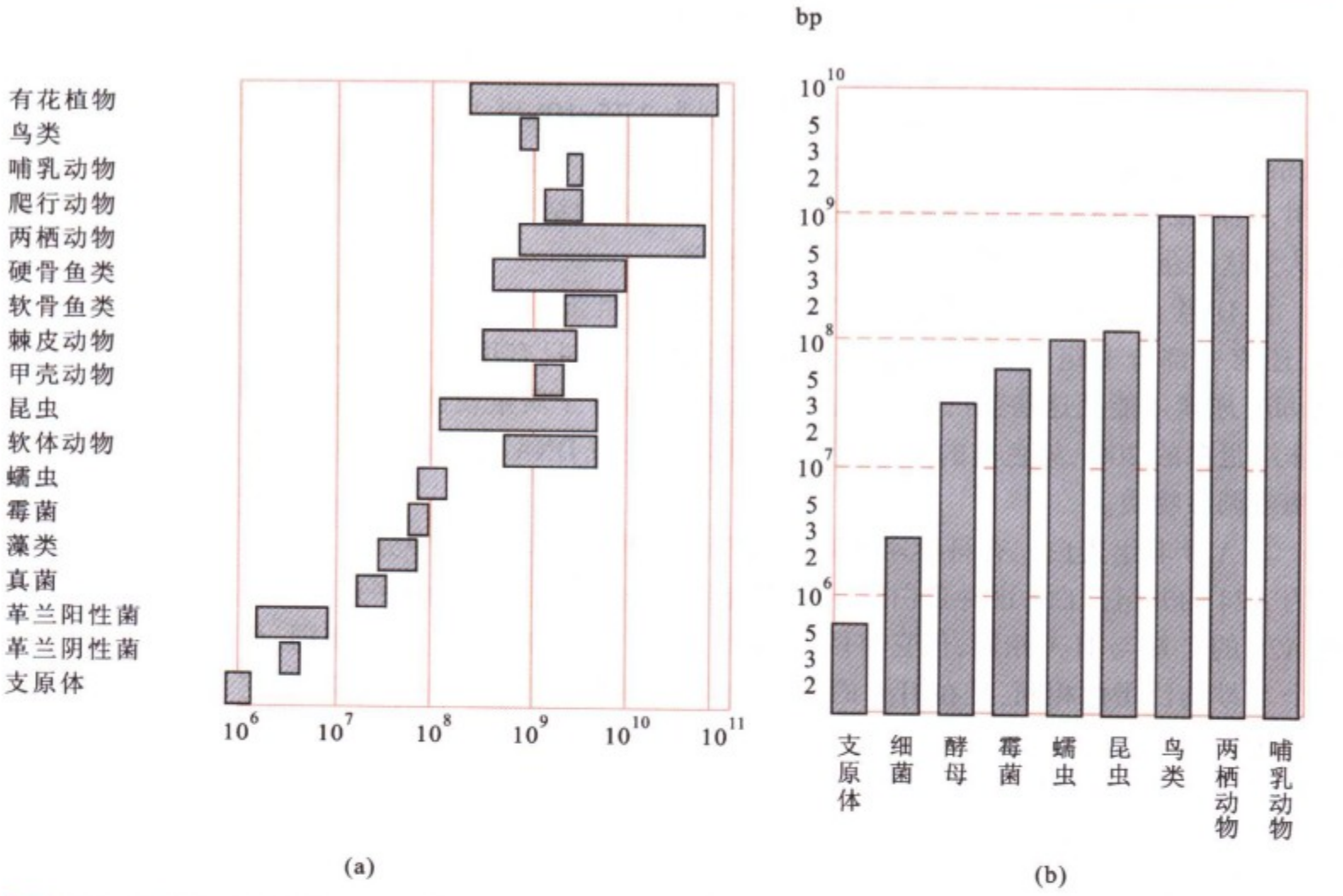



**图 2-5 各种生物细胞内 DNA 总量比较** (a)同类生物不同种属之间 DNA总量变化很大;(b)从编码每类生物所需要 DNA 量的最低值看,生物细胞中的C 值具有从低等生物到高等生物逐渐增加的趋势。

重要的。

(2) 中度重复序列 这类序列的重复次数在 10~$10^4$, 占总 DNA 的 10%~40%,如在小鼠中占 20%,在果蝇中占 15%,各种 rRNA、tRNA 以及某些结构基因如组蛋白基因等都属于这一类。

28S、5.8S 及 18S rRNA 基因: 非洲爪蟾的 18S、5.8S 及 28S rRNA 基因是连在一起的,它们中间隔着不转录的间隔区,这些 18S、5.8S 及 28S rRNA 基因及间隔区组成的单位在 DNA 链上串联重复约 5 000 次(图 2-6),用原位杂交方法证实这些基因位于染色体的核仁组织者区。不转录的间隔区是由 21~100 bp 组成的类似卫星 DNA 的串联重复序列。在许多动物的卵细胞形成过程中这些基因可进行几千次不同比例的复制,产生 $2\times10^6$ 个拷贝,使 rDNA 占卵细胞 DNA 的 75%,从而使该细胞能积累 $10^{12}$ 个核糖体,以合成大量蛋白质供细胞分裂之需。若无这种放大机制,可能要经历几个世纪才能积累 $10^{12}$ 个核糖体。中度重复

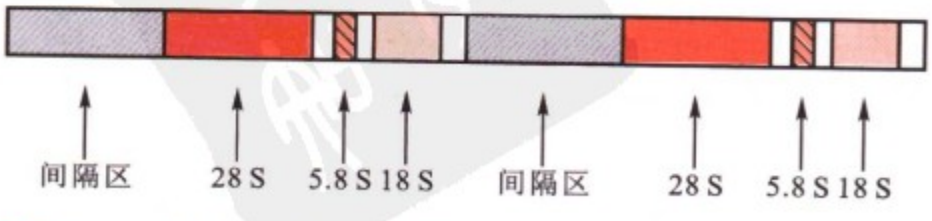



**图 2-6 非洲爪蟾的 rRNA 基因结构示意图**

序列往往分散在不重复序列之间。

(3) 高度重复序列——卫星 DNA　这类 DNA 只在真核生物中发现,占基因组的 10%~60%,由 6~100 个碱基组成,在 DNA 链上串联重复成千上万次。实验中常用 CsCl 密度梯度离心将卫星 DNA 与其他 DNA 分开, 形成两个以上的峰,即含量较大的主峰和高度重复序列小峰,后者又称卫星区带(峰)。果蝇 DNA 在中性 CsCl 密度梯度离心中出现 1 条主带和 3 条卫星区带,人 DNA 经 CsCl 离心后可得到 4 条卫星区带。原位杂交法证明,许多卫星 DNA 均位于染色体的着丝粒,也有一些在染色体臂上。这类 DNA 是高度浓缩的,是异染色质的组成部分。卫星 DNA 是不转录的,其功能不明,可能与染色体的稳定性有关。

3. 染色质和核小体

由 DNA 和组蛋白组成的染色质纤维细丝是许多核小体连成的念珠状结构,大量实验证实了这一结构模型。首先,人们发现染色质 DNA 的 $T_m$ 值比自由 DNA 高,说明在染色质中 DNA 极可能与蛋白质分子相互作用;其次,在染色质状态下, 由 DNA 聚合酶和 RNA 聚合酶催化的 DNA 复制和转录活性大大低于在自由DNA 中的反应;DNA 酶Ⅰ(DNase Ⅰ)对染色质 DNA 的消化远远慢于对纯 DNA 的消化。另外,染色质的电子显微镜图显示出由核小体组成的念珠状结构,可以看到由一条细丝连接着的一连串直径为 10 nm 的球状体。在染色质的 X 线衍射图中也发现了 10 nm 的重复单位。DNaseI 或小球菌核酸酶可以切断游离 DNA 上的任何磷酸二酯键,而除了少数几个位点以外染色质中的 DNA 是受保护不被消化的。用小球菌核酸酶处理染色质以后进行电泳,便可以得到一系列片段,这些被保留的 DNA 片段均为 200 bp 基本单位的倍数,其大小分别为 200 bp(单体)、400 bp(双体)、600 bp(三体)等。电镜图显示一个染色质片段中球形颗粒的长度为 200 bp,如具有 600 bp DNA 的染色质片段含有 3 个直径为 10 nm 的颗粒。因此,电镜上看到的颗粒相当于核酸酶切割后得到的核小体(图 2-7)。

现在已经知道,核小体是由 $H_2A$、$H_2B$、$H_3$、$H_4$ 各两个分子生成的八聚体和由大约 200 bp 的 DNA 组成的。八聚体在中间,DNA 分子盘绕在外,而 $H_1$ 则在核小体的外面。每个核小体只有一个 $H_1$。所以,核小体中组蛋白和 DNA 的比例是每 200 bp DNA 有 $H_2A$、$H_2B$、$H_3$、$H_4$ 各两个,$H_1$ 一个。用核酸酶水解核小体后产生只含 146 bp 核心颗粒 (图 2-7), 包括组蛋白八聚体及与其结合的 146 bp DNA,该序列绕在核心外面形成 1.75 圈,每圈约 80 bp。由许多核小体构成了连续的染色质 DNA 细丝。

核小体的形成是染色体中 DNA 压缩的第一个阶段。在核小体中 DNA 盘绕组蛋白八聚体核心,从而使分子收缩至原尺寸的 1/7。200 bp DNA 完全舒展时长约 68 nm,却被压缩在 10 nm 的核小体中。当然 1/7 的压缩比仍远远低于染色体中 DNA 的压缩比。例如,人中期染色体中含 $6.2\times10^9$ bp,其理论长度应是 200 cm,这么长的 DNA 被包装在 46 个 5 μm 长的圆柱体(染色体)中,其压缩比约为$10^4$。分裂间期染色质比较松散,压缩比大约是 $10^2$~$10^3$。所以说,核小体只是

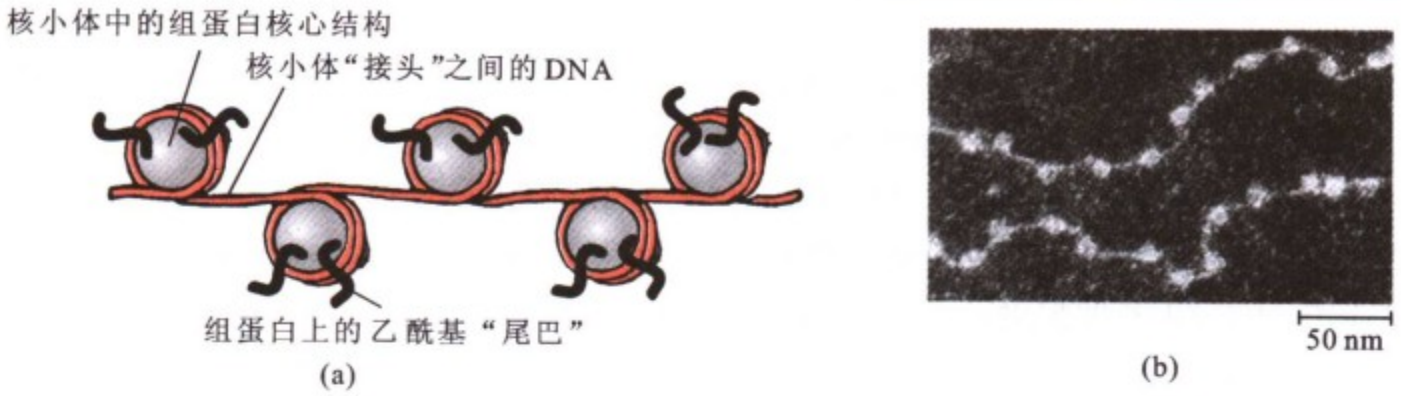



**图 2-7 核小体单位的存在及核心颗粒的形成** (a)核小体结构示意图;(b)从电镜下看到的染色质结构;(c)核小体单元的产生。

DNA 压缩的第一步。

那么,DNA 包装成染色体的下一个水平是什么呢? 在电镜下观察染色质,可以看到 10 nm 及 30 nm 两类不同的纤维构造。通过改变离子强度,30 nm 纤维和 10 nm 纤维可以可逆地相互转化。10 nm 纤维是由核小体串联成的染色质细丝,主要在低离子强度及无 $H_1$ 情况下产生。当离子强度较高且有 $H_1$ 存在时,以30 nm 纤维为主,它是由 10 nm 的染色质细丝盘绕成螺旋管状

的粗丝，通称螺线管(solenoid)。螺线管的每一螺旋包含 6 个核小体，其压缩比为 6。这种螺线管是分裂间期染色质和分裂中期染色体的基本组分。染色质和染色体可能是一个多层次的螺旋结构，上述螺线管可进一步压缩形成超螺旋。有迹象表明：中期染色质是一细长、中空的圆筒，直径为 4 000 nm，由 30 nm 螺线管缠绕而成，压缩比是 40。这个超螺旋圆筒进一步压缩 5 倍便成为染色体单体，总压缩比是 7×6×40×5，将近一万倍，与 DNA 的压缩比相近（图 2-8）。表 2-7 总结了染色体形成过程中长度与宽度的变化。

真核生物基因组的结构特点总结归纳如下：

① 真核基因组庞大，一般都远大于原核生物的基因组。

② 真核基因组存在大量的重复序列。

③ 真核基因组的大部分为非编码序列，占整个基因组序列的 90%以上，该特点是真核生物与细菌和病毒之间最主要的区别。

④ 真核基因组的转录产物为单顺反子。

⑤ 真核基因是断裂基因，有内含子结构。

⑥ 真核基因组存在大量的顺式作用元件。包括启动子、增强子、沉默子等(详见“8 真核基因表达调控”)。

⑦ 真核基因组中存在大量的 DNA 多态性。DNA 多态性是指 DNA 序列中发生变异而导致的个体间核苷酸序列的差异，主要包括单核苷酸多态性(single nucleotide polymorphism，SNP)和串联重复序列多态性（tandem repeats polymorphism）两类。

⑧ 真核基因组具有端粒结构。端粒(telomere）是真核生物线性基因组 DNA 末端的一种特殊结构，它是一段 DNA 序列和蛋白质形成的复合体。其 DNA 序列相当保守，一般由多个短寡核苷酸串联在一起构

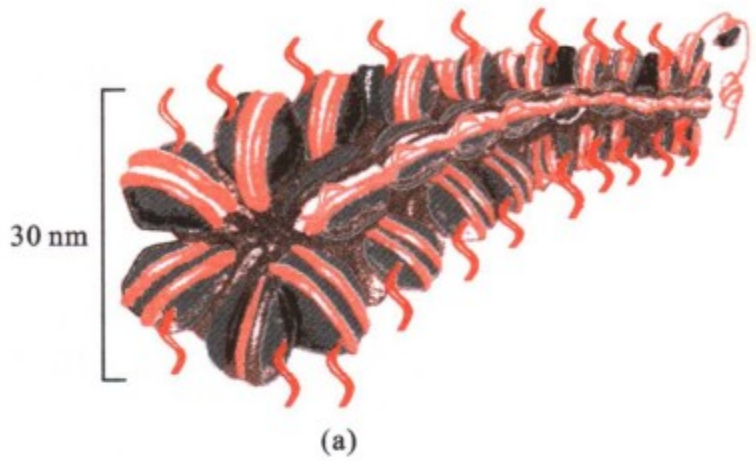



**图 2-8 染色体 DNA 结构示意图** (a)30 nm 染色体结构；(b)双螺旋 DNA 经过一系列包装成为染色体的过程。

表 2-7 染色体形成过程中长度与宽度的变化

| 过程 | 成分 | 名称 | 宽度增加 | 长度压缩 |
|---|---|---|---|---|
| 第一级 | DNA + 组蛋白 | 核小体 | 5 倍 | 7 倍 |
| 第二级 | 核小体 | 螺线管 | 3 倍 | 6 倍 |
| 第三级 | 螺线管 | 超螺旋 | 13 倍 | 40 倍 |
| 第四级 | 超螺旋 | 染色体 | 2.5 ~ 5 倍 | 5 倍 |
| 合计 | | | 500 ~ 1 000 倍 | 8 400 倍<br>(8 000 ~ 10 000) |

成。人类的端粒 DNA 长约5~15 kb。它具有保护线性 DNA 的完整复制、保护染色体末端和决定细胞的寿命等功能,有关端粒的研究也是分子生物学的研究热点之一。

### 2.1.3 原核生物基因组

原核生物的基因组很小,大多只有一条染色体,且 DNA 含量少,如大肠杆菌 DNA 的相对分子质量仅为 $2.4\times10^9$, 或 $4.6\times10^6$ bp, 其完全伸展总长约为1.3 mm,含 4 000 多个基因。最小的病毒如双链 DNA 病毒 SV40,其基因组相对分子质量只有 $3\times10^6$,含 5 个基因;而单链 RNA 病毒 Qβ,只含有 4 个基因。此外,细菌的质粒、真核生物的线粒体、高等植物的叶绿体等也含有 DNA 和功能基因,这些 DNA 被称为染色体外遗传因子。从基因组的组织结构来看,原核细胞 DNA 有如下特点。

1. 结构简练

原核 DNA 分子的绝大部分是用来编码蛋白质的, 只有非常小的一部分不转录, 这与真核 DNA 的冗余现象不同。在 ΦX174 中不转录部分只占 4%左右(217/5 386),T4 DNA 中占 5.1%(282/5 577),而且,这些不转录 DNA 序列通常是控制基因表达的序列,如 ΦX174 的 *H* 和 *A* 基因之间(3 906~3 973 位核苷酸)就包括了 RNA 聚合酶结合位点、转录终止信号区及核糖体结合位点等基因表达调控元件。

2. 存在转录单元

原核生物 DNA 序列中功能相关的 RNA 和蛋白质基因, 往往丛集在基因组的一个或几个特定部位,形成功能单位或转录单元,它们可被一起转录为含多个 mRNA 的分子,称为多顺反子 mRNA。ΦX174 及 G4 基因组中就含有数个多顺反子。功能相关的基因,如 ΦX174 中的 D-E-J-F-G-H 等都串联在一起转录产生一条 mRNA 链,然后再翻译成各种蛋白质,其中 J、F、G 及 H 编码外壳蛋白,D 蛋白与病毒装配有关,E 蛋白则导致细菌的裂解。这是功能相关基因协同表达的方式之一。在大肠杆菌中,由几个结构基因及其操纵基因、启动基因组成的操纵子也是一种转录功能单位,如组氨酸操纵子转录成一条多顺反子 mRNA, 再翻译成组氨酸合成途径中的9个酶。

3. 有重叠基因

直到不久前，人们还认为基因是一段 DNA 序列，这段序列负责编码一个蛋白质或一条多肽。但是，已经发现在一些细菌和动物病毒中有重叠基因，即同一段 DNA 能携带两种不同蛋白质的信息。1973 年 Weiner 和 Weber 在研究一种大肠杆菌 RNA 病毒时发现，有两个基因从同一起点开始翻译，一个在 400 bp 处结束，生成较小的蛋白质；而在 3%的情况下，翻译可一直进行下去直到 800 bp 处碰到双重终止信号时才停止，合成相对分子质量较大的蛋白质。当时他们认为相对分子质量大的蛋白质含量少，对病毒无关紧要，因而不予重视，没有进一步研究。后来，Weissman 证实小分子蛋白质是一种外壳蛋白，需要量大，大分子蛋白质产量虽少，却是组成有感染力的病毒颗粒所必需的。当 Weiner 等想回头研究这一现象时，Sanger 在《自然》杂志(1977 年)上发表了 ΦX174 DNA 的全部核苷酸序列，正式发现了重叠基因。

ΦX174 是一种单链 DNA 病毒，宿主为大肠杆菌，感染宿主后合成另一条链[(-)链]，变成复制型(replicating form，RF)，然后以新合成的(-)链为模板合成子代 DNA 分子[(+)链]，并合成 9 个蛋白质，总相对分子质量约 $2.5\times10^5$，相当于6 078 个核苷酸，而病毒 DNA 本身只有 5 375 个核苷酸，顶多能编码总相对分子质量为 $2.0\times10^5$ 的多肽，这个矛盾很长时期都无法解决。Sanger 在弄清 ΦX174 DNA 的全部核苷酸序列及各个基因的起讫位置和密码数目以后发现，ΦX174 的 9 个基因有些是重叠的。主要有以下几种情况：

① 一个基因完全在另一个基因里面，如基因 *B* 在基因 *A* 内，基因 *E* 在基因 *D* 内。

② 部分重叠，如基因 *K* 和基因 *C* 是部分重叠的。

③ 两个基因只有一个碱基对是重叠的，如 *D* 基因终止密码子的最后一个碱基是 *J* 基因起始密码子的第一个碱基。

尽管这些重叠基因的 DNA 序列大致相同，但由于基因重叠部位一个碱基的变化可能影响后续肽链的全部序列，从而编码完全不同的蛋白质。除了 ΦX174 外，SV40 病毒、G4 噬菌体的 DNA 中也存在基因重叠现象。如 SV40 DNA 由 5 224 个碱基对组成，它编码 3 个外壳蛋白(VP1、VP2、VP3)及 2 个表面抗原(T 及 t)。Fiers 等在测定 SV40 DNA 的全部核苷酸顺序以后发现，*VP1*、*VP2*、*VP3* 基因之间都有122 个碱基对的重叠序列，但密码子各不相同。t 抗原基因完全在 T 抗原基因里面，它们有一个共同的起始密码子。基因重叠可能是生物进化过程中自然选择的结果。

## 2.2 DNA 的结构

现在我们已经知道，DNA 是遗传的物质基础，基因是具有特定生物功能的 DNA 序列，通过基因的表达能够使上一代的性状准确地在下一代表现出来。那么，DNA 为什么能起遗传作用?它又是怎样起作用的呢?这与它的分子结构是密切相关的。

## 2.2.1 DNA 的一级结构

Watson 和 Crick 于 1953 年提出了著名的 DNA 双螺旋模型，为合理地解释遗传物质的各种功能，解释生物的遗传和变异，解释自然界千变万化的生命现象奠定了理论基础。前面说过，DNA 又称脱氧核糖核酸，是 deoxyribonucleic acid 的缩写。它是一种高分子化合物，其基本单位是脱氧核苷酸。在所有的 DNA 分子中，磷酸和脱氧核糖(图 2-9)是永远不变的，而含氮碱基是可变的，主要有 4 种，即腺嘌呤(A)、鸟嘌呤(G)、胞嘧啶(C)和胸腺嘧啶(T)(表 2-8)。因此脱氧核

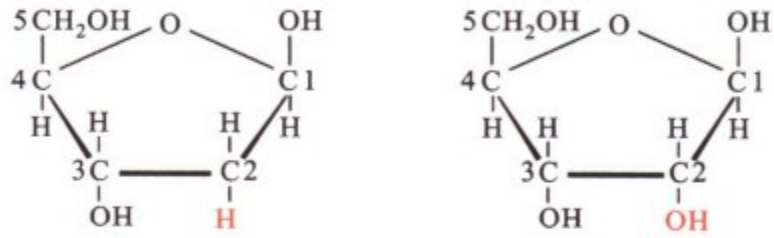



图 2-9 2-脱氧核糖(左)和核糖(右)结构式

表 2-8 碱基、核苷和核苷酸

| 碱 基 | 核 苷 | 核苷酸 | RNA | DNA |
|---|---|---|---|---|
| 腺嘌呤(adenine) | 腺苷(adenosine) | 腺苷酸(adenylic acid) | AMP | dAMP |
| 鸟嘌呤(guanine) | 鸟苷(guanosine) | 鸟苷酸(guanylic acid) | GMP | dGMP |
| 胞嘧啶(cytosine) | 胞苷(cytidine) | 胞苷酸(cytidylic acid) | CMP | dCMP |
| 胸腺嘧啶(thymine) | 胸苷(thymidine) | 胸苷酸(thymidylic acid) | | dTMP |
| 尿嘧啶(uracil) | 尿苷(uridine) | 尿苷酸(uridylic acid) | UMP | |

苷酸可分别称为腺嘌呤脱氧核苷酸、鸟嘌呤脱氧核苷酸、胞嘧啶脱氧核苷酸和胸腺嘧啶脱氧核苷酸。许许多多个脱氧核苷酸经 3′→5′磷酸二酯键聚合而成为 DNA 链(图 2-10)。DNA 通常以线性或环状形式存在，绝大多数 DNA 分子都由两条碱基互补的单链构成，只有少数生物，如某些噬菌体或病毒就是以单链形式存在的。

所谓 DNA 的一级结构，就是指4 种核苷酸的连接及排列顺序，表示了该 DNA 分子的化学构成。核苷酸序列对DNA 高级结构的形成有很大影响，如 B-DNA 中多聚 G-C 区易形成左手螺旋DNA(Z-DNA)，而反向重复的 DNA 片段易出现发卡式结构等。DNA 不仅具有严格的化学组成，还具有特殊的空间结构，它主要以有规则的双螺旋形式存在，其基本特点是：

① DNA 分子是由两条互相平行的脱氧核苷酸长链盘绕而成的。

② DNA 分子中的脱氧核糖和磷酸交替连接，排在外侧，构成基本骨架，碱基排列在内侧。

③ 两条链上的碱基通过氢键相结合，形成碱基对，它的组成有一定的规律。这就是嘌呤与嘧啶配对，而且腺嘌呤(A)只能与胸腺嘧啶(T)配对，鸟嘌呤(G)只能与胞嘧啶(C)配对。如一条链上某一碱基是 C，另一条链上与它配对的

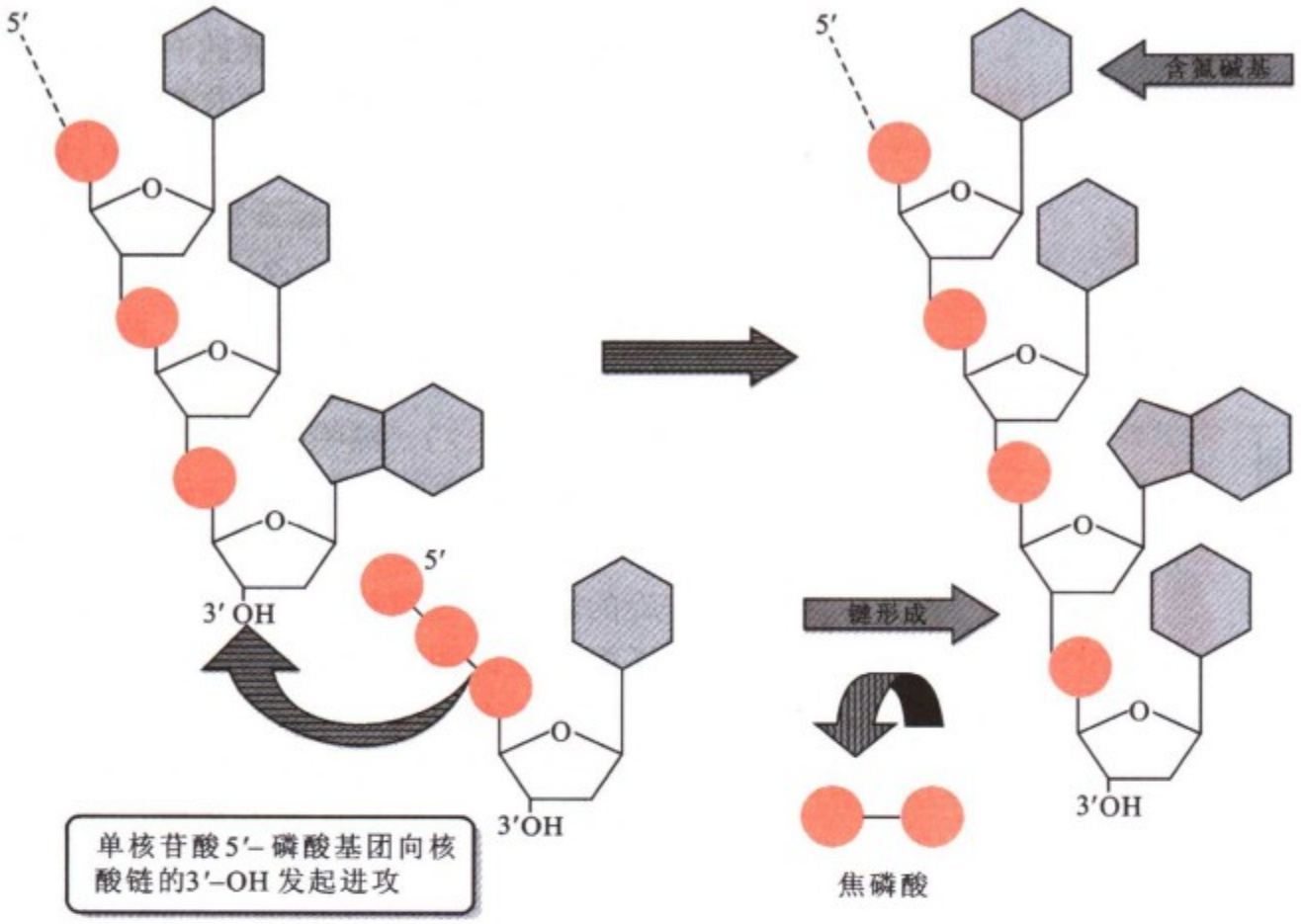



图 2-10 多聚核苷酸链中新生磷酸糖苷键的产生过程示意图

碱基必定是 G。碱基之间的这种一一对应的关系叫碱基互补配对原则。从 DNA 的分子结构可以看出,碱基在长链中的排列顺序是千变万化的。组成 DNA 分子的碱基虽然只有 4 种,它们的配对方式也只有 A 与 T,C 与 G 两种,但是,由于碱基可以任何顺序排列,构成了 DNA 分子的多样性。例如,某 DNA 分子的一条多核苷酸链由 100 个不同的碱基组成,它的可能排列方式就是 $4^{100}$。实际上,每条多核苷酸链中的碱基总数远远超过 100 个,所以,它们的排列方式几乎是无限的。每个 DNA 分子所具有的特定的碱基排列顺序构成了 DNA 分子的特异性,不同的 DNA 链可以编码出完全不同的多肽。

### 2.2.2 DNA 的二级结构

DNA 的二级结构是指两条多核苷酸链反向平行盘绕所生成的双螺旋结构。在生物活体中,不论 DNA 的二级结构还是高级结构,都是时刻变化的,即在二级结构的各种构象间、在二级结构与高级结构间或高级结构的各种构象间存在一个动力学的平衡。通常情况下,DNA 的二级结构分两大类:一类是右手螺旋,如 A-DNA 和 B-DNA,DNA 通常是以右手螺旋形式存在的;另一类是左手螺旋,即 Z-DNA(图 2-11)。

1. DNA 右手螺旋的几种构象及其动态平衡

Watson 和 Crick 由 X 线衍射结构分析提出了 DNA 分子双螺旋结构模型,此模型所描述的是 B-DNA 钠盐在一定湿度下的结构。B-DNA 钠盐结构既规则

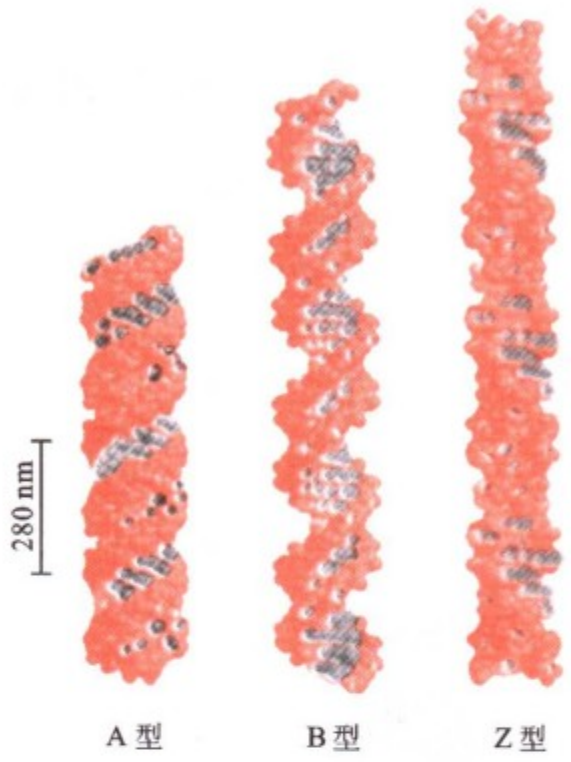



图 2-11 A-,B-,Z-型 DNA 结构比较

又很稳定，是由两条反向平行的多核苷酸链围绕同一中心轴构成的右手螺旋结构。多核苷酸的方向由核苷酸间的磷酸二酯键的走向决定，一条从 5′→3′，另一条从 3′→5′。链间有螺旋形的凹槽，其中一个较浅，叫小沟（约 1.2 nm 交叉）；一个较深，叫大沟（约 2.2 nm 交叉）。两条链上的碱基以氢键相连，G 与 C 配对，A 与 T 配对（图 2-12）。嘌呤和嘧啶碱基对层叠于双螺旋的内侧。顺着螺旋轴心从上向下看，可见碱基平面与纵轴垂直，且螺旋的轴心穿过氢键的中点。相邻碱基对平面之间的距离为 0.34 nm，即顺中心轴方向，每隔 0.34 nm 有一个核苷酸，以3.4 nm 为一个结构重复周期。核苷酸的磷酸基团与脱氧核糖在外侧，通过磷

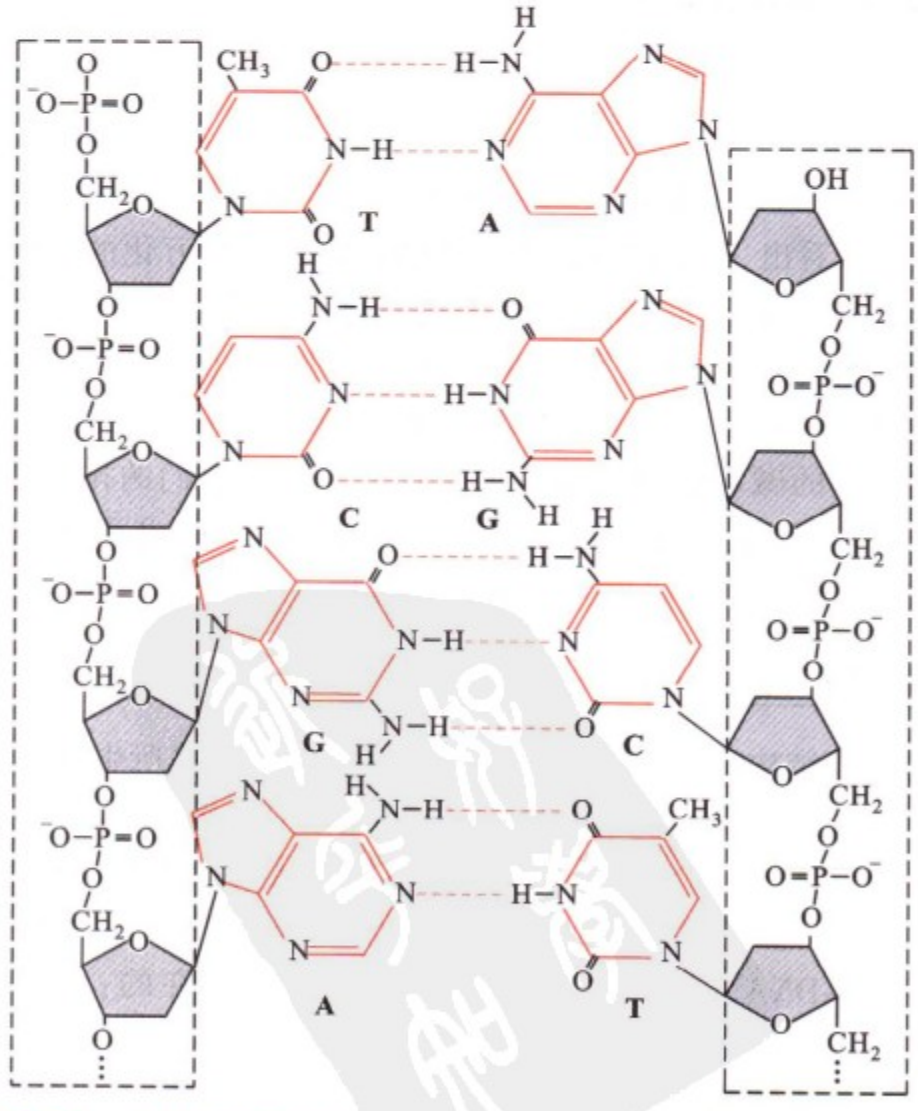



图 2-12 DNA 双螺旋结构中碱基配对示意图（A 与 T，G 与 C 配对）

酸二酯键相连接而构成 DNA 分子的骨架。脱氧核糖环平面与纵轴大致平行。双螺旋的直径为 2.0 nm(图 2-13)。

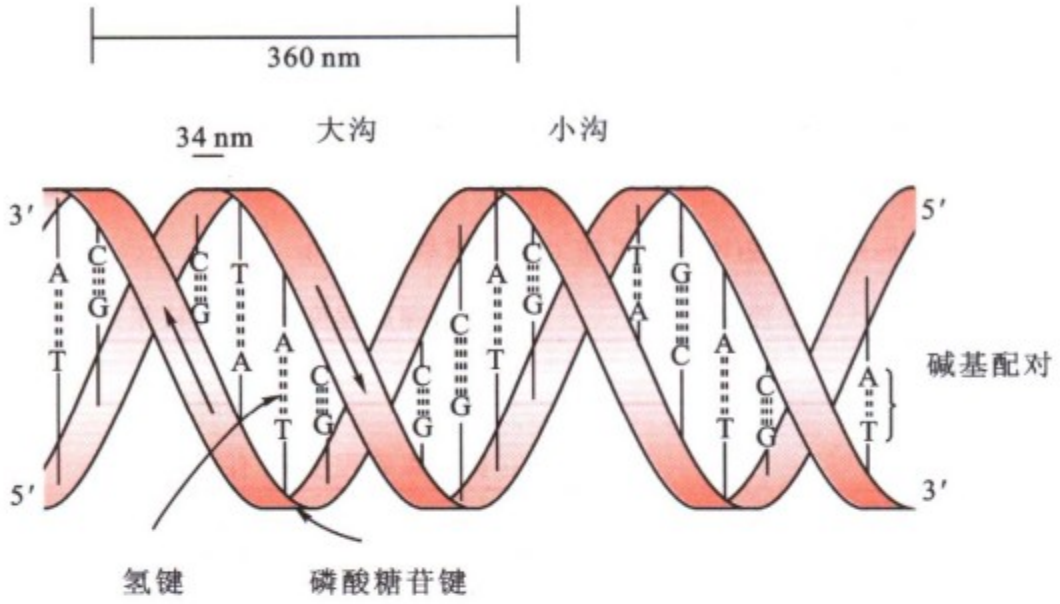



图 2-13 DNA 的反向平行双螺旋结构

一般说来,A-T 丰富的 DNA 片段常呈 B-DNA, 它是普遍存在的结构。若 DNA 双链中一条链被相应的 RNA 链所替换,则变构成 A-DNA。当 DNA 处于转录状态时,DNA 模板链与由它转录所得的RNA 链间形成的双链就是 A-DNA。由此可见 A-DNA 构象对基因表达有重要意义。此外,B-DNA 双链都被 RNA 链所取代而得到由两条 RNA 链组成的双螺旋结构也是 A-DNA,表 2-9 列举了几种主要 DNA 的结构参数。除 A-DNA、B-DNA 螺旋外,还存在 B′-DNA、C-DNA 和 D-DNA 等不同形式。

表 2-9 不同螺旋形式 DNA 分子主要参数比较

| 双螺旋 | 碱基倾角(°) | 碱基间距(nm) | 螺旋直径(nm) | 每轮碱基数 | 螺旋方向 |
|---|---|---|---|---|---|
| A - DNA | 20 | 0.26 | 2.6 | 11 | 右 |
| B - DNA | 6 | 0.34 | 2.0 | 10 | 右 |
| Z - DNA | 7 | 0.37 | 1.8 | 12 | 左 |

2. DNA 二级结构中左手螺旋——Z-DNA 的研究

Z-DNA 结构是 1979 年由 Rich 提出的。Rich 用一位荷兰科学家提供的 d(CGCGCG)结晶,对它进行 X 线衍射结构分析,最后提出了 Z-DNA 结构模型。当时, 由于用右手螺旋结构模型解链过程来解释大分子 DNA 复制时遇到的困难,该模型的提出曾一度动摇过右手螺旋学说,现已证明,左手螺旋 Z-DNA 只是右手螺旋结构模型的一个补充和发展。

虽然 B-DNA 是最常见的 DNA 构象, 但 A-DNA 和 Z-DNA 似乎具有不同的生物活性。图 2-14 是 Z-DNA 调控基因转录的两个模式。在邻近调控系统中,与调节区相邻的转录区被 Z-DNA 抑制, 只有当 Z-DNA 转变为 B-DNA 后,转录才得以活化。而在远距离调控系统中,Z-DNA 可通过改变负超螺旋水平,决

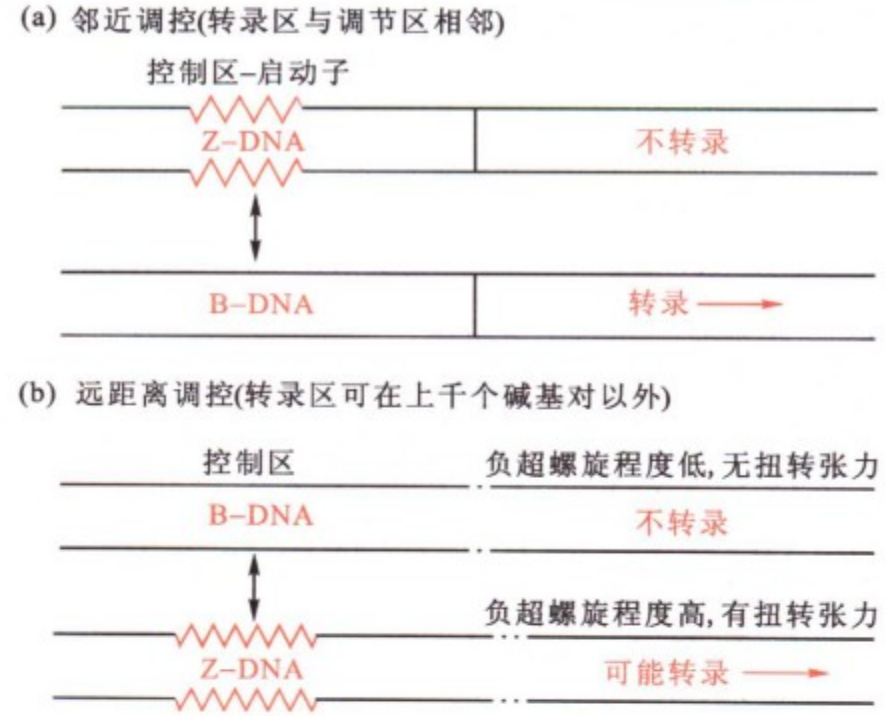



图 2-14 B-DNA 和 Z-DNA 调节转录的示意图

定聚合酶能否与模板链相结合而调节转录起始活性。

### 2.2.3 DNA 的高级结构

DNA 的高级结构是指 DNA 双螺旋进一步扭曲盘绕所形成的特定空间结构。超螺旋结构是 DNA 高级结构的主要形式,可分为正超螺旋与负超螺旋两大类,它们在不同类型的拓扑异构酶作用下或在特殊情况下可以相互转变,如:

$$\text{负超螺旋} \xrightarrow[\text{溴乙啶}]{\text{拓扑异构酶}} \text{松弛 DNA} \xrightarrow[\text{溴乙啶}]{\text{拓扑异构酶}} \text{正超螺旋}$$

图 2-15 用电话机连接做模型说明 DNA 链中的螺旋与超螺旋

DNA 双螺旋结构中,一般每转一圈有 10 个核苷酸对,双螺旋总处于能量最低状态。若正常 DNA 双螺旋额外地多转或少转几圈,使每转一圈的核苷酸数目大于或小于 10,就会出现双螺旋空间结构的改变,在 DNA 分子中产生额外的张力(图 2-15)。若此时双螺旋链的末端是自由的,可通过链的转动而释放这种额外的张力,从而保持原来的双螺旋结构;若此时 DNA 分子的末端是固定的或是环状分子,双链不能自由转动,额外的张力就不能释放而导致 DNA 分子内部原子空间位置的重排,造成扭曲,即出现超螺旋结构(图 2-16)。由此可见,B-DNA 双螺旋分子的链间螺旋数若发生变化,就会出现超螺旋结构,而且超螺旋的绕数与 B-DNA 的链间螺旋数有密切关系。

研究细菌质粒 DNA（环状双链 DNA）时发现，天然状态下该 DNA 以负超螺旋为主，稍被破坏即出现开环结构，两条链均断开则呈线性结构。在电场作用下，相同分子质量的超螺旋 DNA 比线性 DNA 迁移率大，线性 DNA 又比开环的DNA 迁移率大，据此可判断质粒结构是否被破坏。

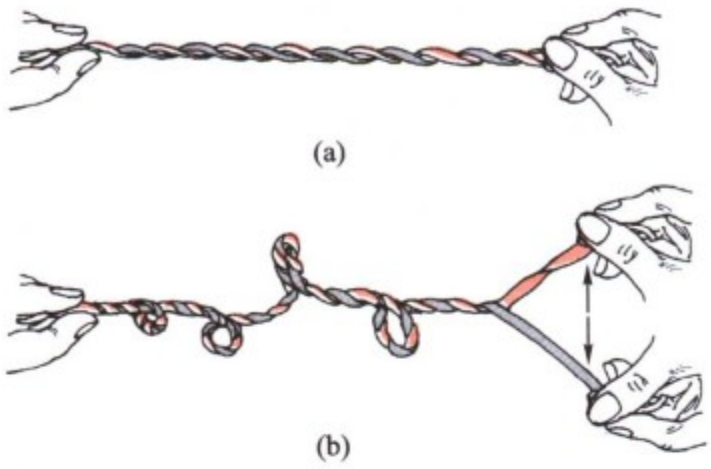



**图 2-16 由于强行分开双螺旋末端而引发产生的超螺旋结构** (a)中间没有超螺旋的两股绳子；(b)一端被强行拉开后，中间出现超螺旋。

DNA 分子的这种变化可以用一个数学公式来表示：

$$L=T+W$$

其中 $L$ 为连接数 (linking number)，是指环形 DNA 分子两条链间交叉的次数。只要不发生链的断裂，$L$ 是个常量。$T$ 为双螺旋的盘绕数(twisting number)，$W$ 为超螺旋数(writhing number)，它们是变量。

如某个 B-DNA 共有 1 350 bp，根据表 2-9，每一螺旋含 10 个碱基对，所以 $L$ 和 $T$ 都是 135，这时 $W$=0，不存在超螺旋。现在若将 DNA 分子的一端固定，另一端朝双螺旋相反方向旋转 5 圈后使两端封闭(此时 $L$=130)，分子内张力可以按两种方式分布，一种在分子内保留一个单链区，即 $T$ 也减为 130，$W$ 仍然为 0。由于 DNA 趋向于保持 B-DNA，因此张力分布的另一种方式是使 $T$ 维持原来的 135，$L$ 既是定值，为了满足上述方程，$W$ 就必须等于-5，该 DNA 分子就成为超盘绕 5 次的负超螺旋。超螺旋是 DNA 三级结构的一种普遍形式，双螺旋 DNA 的松开导致负超螺旋，而拧紧则导致正超螺旋。

## 2.3 DNA 的复制

前面我们讲过，生命的遗传实际上是染色体 DNA 自我复制的结果。而染色体 DNA 的自我复制主要是通过半保留复制来实现的，是一个以亲代 DNA 分子为模板合成子代 DNA 链的过程。细胞分裂时，通过 DNA 准确的自我复制(self-replication)，亲代细胞所含的遗传信息就原原本本地传送到子代细胞。由于 DNA 是遗传信息的载体，因此亲代 DNA 必须以自身分子为模板来合成新的分子——准确地复制成两个拷贝，并分配到两个子代细胞中去，才能真正完成其遗传信息载体的使命。DNA 的双链结构对于维持这类遗传物质的稳定性和复制的准确性都是极为重要的。

双链 DNA 的复制是一个非常复杂的过程，在复制的起始、延伸和终止三个阶段，无论是原核生物还是真核生物都需要有多种酶和蛋白质的协同参与。DNA 复制均涉及拓扑异构酶、解旋酶、单链结合蛋白、引物合成酶、DNA 聚合酶及连接酶等酶和蛋白质的参与。

### 2.3.1 DNA 的半保留复制

Watson 和 Crick 在提出 DNA 双螺旋结构模型时就对 DNA 的复制过程进行了探讨。由于 DNA 分子由两条多核苷酸链组成,两条链上的碱基——G 只能与 C 相配对,A 只能与 T 相配对,所以,两条链是互补的,一条链上的核苷酸排列顺序决定了另一条链上的核苷酸排列顺序。就是说,DNA 分子的每一条链都含有合成它的互补链所需的全部信息。Watson 和 Crick 推测,DNA 在复制过程中碱基间的氢键首先断裂,双螺旋解旋并被分开,每条链分别作为模板合成新链,产生互补的两条链。这样新形成的两个 DNA 分子与原来 DNA 分子的碱基顺序完全一样。因此,每个子代分子的一条链来自亲代 DNA,另一条链则是新合成的, 所以这种复制方式被称为 DNA 的半保留复制 (semiconservative replication)。

1958 年,Meselson 和 Stahl 研究了经 $^{15}N$ 标记 3 个世代的大肠杆菌 DNA,首次证明了 DNA 的半保留复制。他们将大肠杆菌长期在以 $^{15}N$ 作氮源的培养基中培养,得到 $^{15}N$-DNA。由于该 DNA 分子的密度比普通 DNA($^{14}N$-DNA)的密度大,在氯化铯密度梯度离心时,这两种 DNA 形成位置不同的区带。他们用普通培养基(含 $^{14}N$ 的氮源)培养 $^{15}N$ 标记的大肠杆菌,经过一代以后,所有 DNA 的密度都在 $^{15}N$-DNA 和 $^{14}N$-DNA 之间,即形成了一半 $^{15}N$ 和一半 $^{14}N$ 的杂合分子,两代后出现等量的 $^{14}N$ 分子和 $^{14}N$-$^{15}N$ 杂合分子。若再继续培养,可以看到 $^{14}N$-DNA 分子增多,说明 DNA 分子在复制时均可被分成两个亚单位,分别构成子代分子的一半,这些亚单位经过许多代复制仍然保持着完整性。

实验已经证明, 无论是原核生物还是真核生物的 DNA 都是以半保留复制方式遗传的。DNA 的这种半保留复制保证了 DNA 在代谢上的稳定性。经过许多代的复制,DNA 多核苷酸链仍可完整地存在于后代而不被分解掉。这种稳定性与 DNA 的遗传功能相符。

### 2.3.2 复制的起点、方向和速度

复制时,双链 DNA 要解开成两股链分别进行,所以,这个复制起点呈现叉子的形式,被称为复制叉(图 2-17)。许多实验证明,DNA 的复制是由固定的起始点开始的。通常,把生物体的复制单位称为复制子(replicon)。一个复制子只含一个复制起始点。复制叉从复制起点开始沿着 DNA 链连续移动,起始点可以启动单向复制或者双向复制,这主要取决于在复制起点形成一个复制叉还是两个复制叉。在单向复制中,所产生的一个复制叉离开起始点,沿 DNA 链前进;在双向复制中,起始点处产生两个复制叉,它们从起始点开始沿着相反的方向等速前进。

从细菌、酵母、线粒体和叶绿体中鉴定出的起始点的共同特点是含有丰富的 AT 序列,它可能有利于 DNA 复制启动时双链的解开。通常,细菌、病毒和线粒体的 DNA 分子都是作为单个复制子完成复制的, 而真核生物基因组可以同时在多个复制起点上进行双向复制,也就是说它们的基因组包含有多个复制子

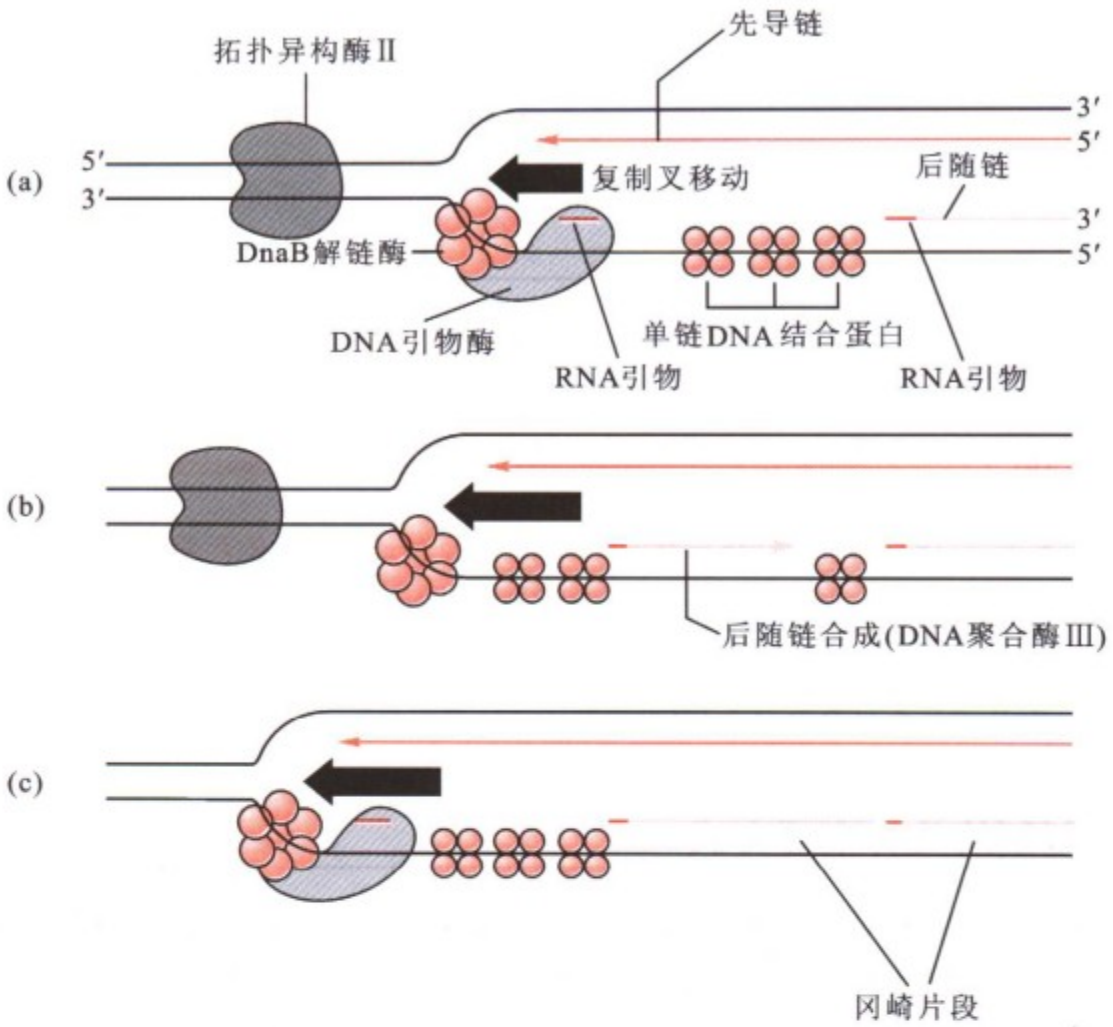



**图 2-17 DNA 链的复制过程示意图** (a)DNA 改变双螺旋构象，解链酶解开双链，在单链 DNA 结合蛋白(SSB)和 DNA 聚合酶Ⅲ的共同作用下合成先导链，方向与复制叉推进的方向一致。在后随链上，当复制叉进一步打开时，RNA 引物才能与 DNA 单链结合；(b)后随链合成时，产生冈崎片段；(c)复制叉继续前进，引物酶合成新的 RNA 引物，与 DNA 单链结合准备引发合成新的冈崎片段。

(表 2-10)。实验表明，无论是原核生物还是真核生物，DNA 的复制主要是从固定的起始点以双向等速方式进行复制的(图 2-18)。复制叉以 DNA 分子上某一特定顺序为起点，向两个方向等速生长前进。

归纳起来，对一个生物体基因组而言，复制起点是固定的，表现为固定的序列，并识别参与复制起始的特殊蛋白质。复制叉移动的方向和速度虽是多种多样的，但以双向等速方式为主。

**表 2-10 部分生物复制子的比较**

| 物 种 | 细胞内复制子数目/个 | 平均长度/kb | 复制子移动速度/(bp · min$^{-1}$) |
|---|---|---|---|
| 大肠杆菌 | 1 | 4 200 | 50 000 |
| 酵母 | 500 | 40 | 3 600 |
| 果蝇 | 3 500 | 40 | 2 600 |
| 爪蟾 | 15 000 | 200 | 500 |
| 蚕豆 | 35 000 | 300 | ? |

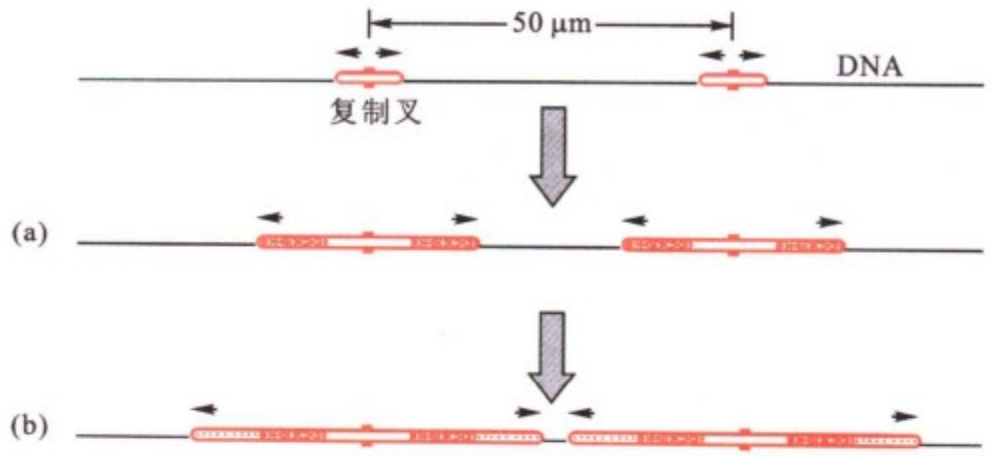



**图 2-18 放射性标记实验证明 DNA 的复制是从固定的起始点双向等速进行的** (a)仅用 $^3$H-胸腺嘧啶标记 DNA 10 min 即压 X 线片；(b)$^3$H 标记 DNA 合成 10 min 后，又继续非标记 DNA 合成 10 min，导致 X 线片上的复制叉变长，银颗粒分布广而稀少。

### 2.3.3 复制的几种主要方式

生物体内 DNA 双链的复制大都是以半保留方式进行的，这是共性；但不同生物体内 DNA 分子的存在形式各不相同(有大、有小、有线性分子、有环状分子)，功能状态也不一样(有的保守、有的极为活跃)，因而反映在复制方式上也有差别，这是个性。绝大多数复制是以复制叉的形式进行的，下面总结了几种复制的机制。

1. 线性 DNA 双链的复制

复制叉生长方式有单一起点的单向(如腺病毒)及双向(如 T7 噬菌体)和多个起始点的双向几种。DNA 双向复制时，复制叉处呈“眼”型(图 2-18)，因为已知的 DNA 聚合酶和 RNA 聚合酶都只能从 5′端向 3′端移动。线性 DNA 在复制中，当 RNA 引物被切除后，留下 5′端的部分单链 DNA，不能为 DNA 聚合酶所作用，使子链短于母链，因此，线性 DNA 复制子末端的复制需要有下列特殊的机制：

(1) 将线性复制子转变为环状或多聚分子。T4 和 T7 噬菌体就是利用这种机制。T4 噬菌体的 DNA 通过其末端的简并性使不同链的 3′端因互补而结合，其缺口被聚合酶作用填满，再经 DNA 连接酶作用生成二联体。这个过程可重复进行直到生成原长 20 多倍的多联体，并由噬菌体 DNA 编码的核酸酶特异切割形成单位长度的 DNA 分子。

(2) 在 DNA 末端形成发夹式结构，使该分子没有游离的末端。草履虫的线性线粒体 DNA 就是一个典型的例子。

(3) 在某种蛋白质(如末端蛋白，terminal protein)的介入下，在真正的末端上启动复制。Φ29 噬菌体 DNA 和腺病毒 DNA 的复制主要依赖这一机制。如图 2-19 所示，这类 DNA 依靠链取代法，从一个末端启动一条新链合成，以此取代原来在双链中配对的 DNA 链。当复制叉到达分子的另一末端时，被替换的链会

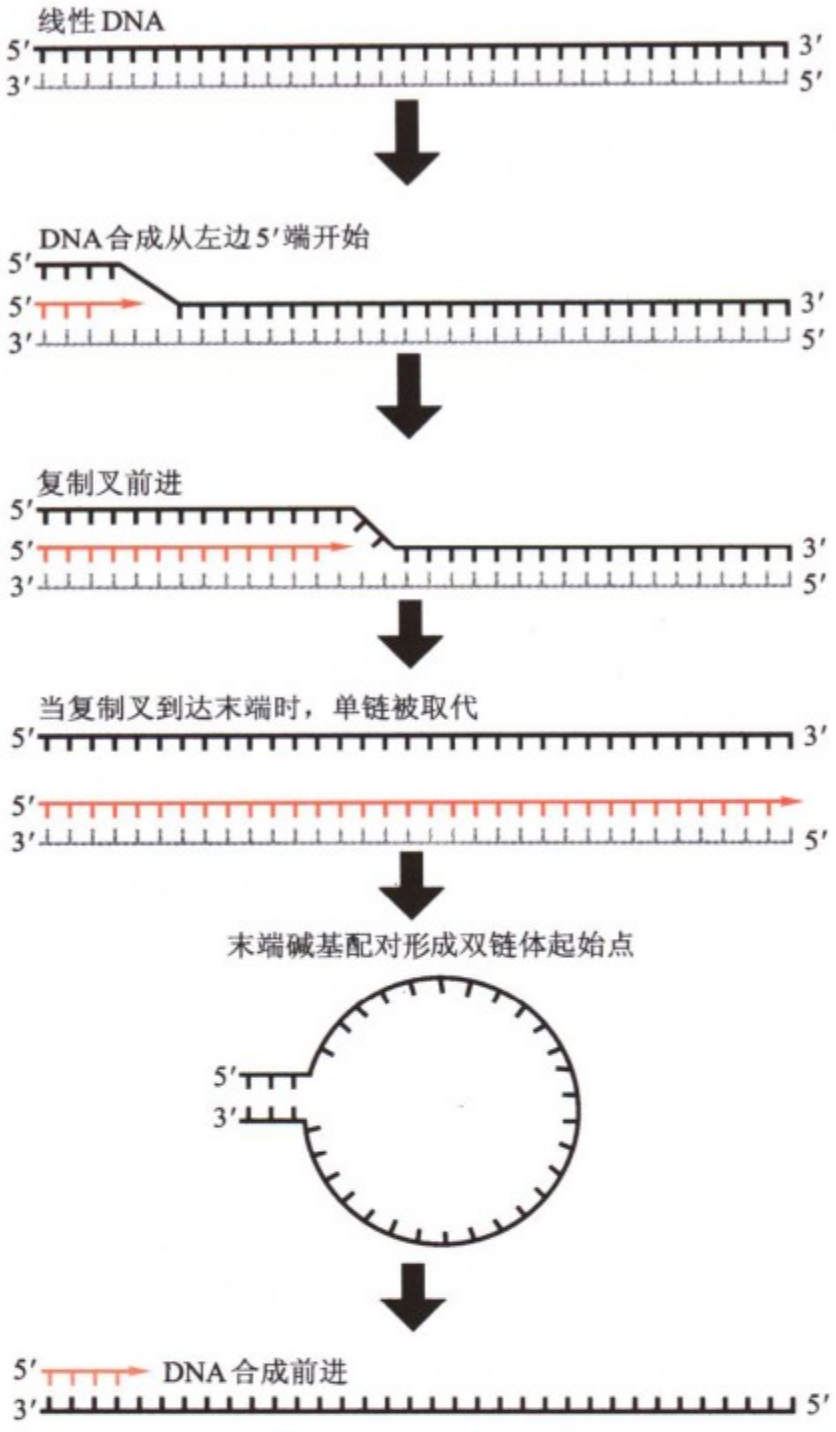



图 2-19 腺病毒 DNA 分别在分子两端启动复制，并通过链取代法使之延伸

作为一条游离的单链而被释放。接着，它又被独立地复制。复制时，5′端首先与末端蛋白共价结合，并由末端蛋白提供一个 C 端来引发一条新的 DNA 链的合成。

2. 环状 DNA 双链的复制

环状双链 DNA 的复制可分为 θ 型、滚环型和 D-环型几种类型。

(1) θ 型　大肠杆菌基因组的复制原点位于天冬酰胺合酶和 ATP 合酶操纵子之间，全长 245 bp，称为 OriC。它富含 AT，并含有多个短的重复序列，能够被复制起始点结合蛋白所识别。复制的起始点涉及 DNA 双链的解旋和松开，形成两个方向相反的复制叉(图 2-20)。前导链 DNA 开始复制前，复制原点的核酸序列被转录生成短 RNA 链，作为起始 DNA 复制的引物。

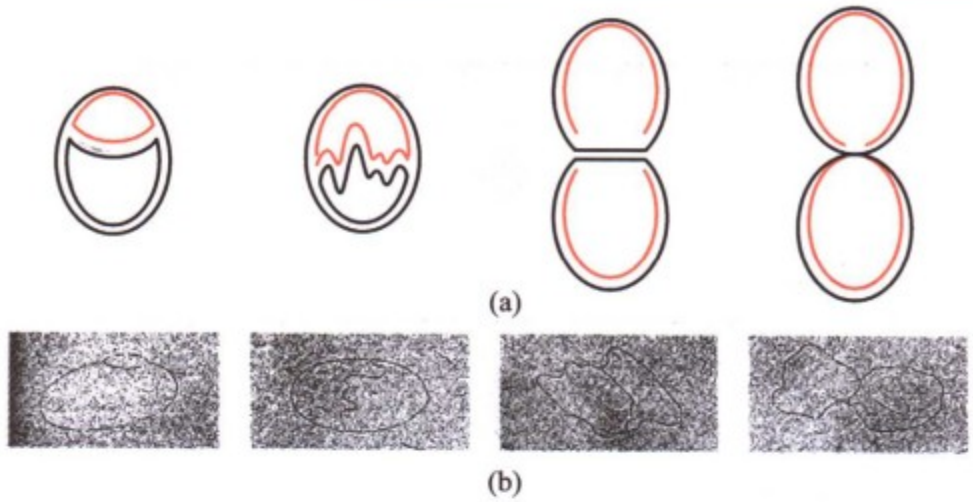



**图 2–20 大肠杆菌质粒 DNA 双向复制的模式与实验验证** (a)θ 形复制模式图;(b)$^3$H 标记质粒 DNA 合成图。

(2) 滚环型(rolling circle) 这是单向复制的一种特殊方式。滚环复制在噬菌体中是很常见的,如 ΦX174 的双链环状 DNA 复制型(RF)就是以这种方式复制的。DNA 的合成由对正链原点的专一性切割开始,所形成的自由 5′ 端被从双链环中置换出来并为单链 DNA 结合蛋白所覆盖,使其 3′–OH 端在 DNA 聚合酶的作用下不断延伸。在这个过程中,单链尾巴的延伸与双链 DNA 的绕轴旋转同步(图 2–21)。

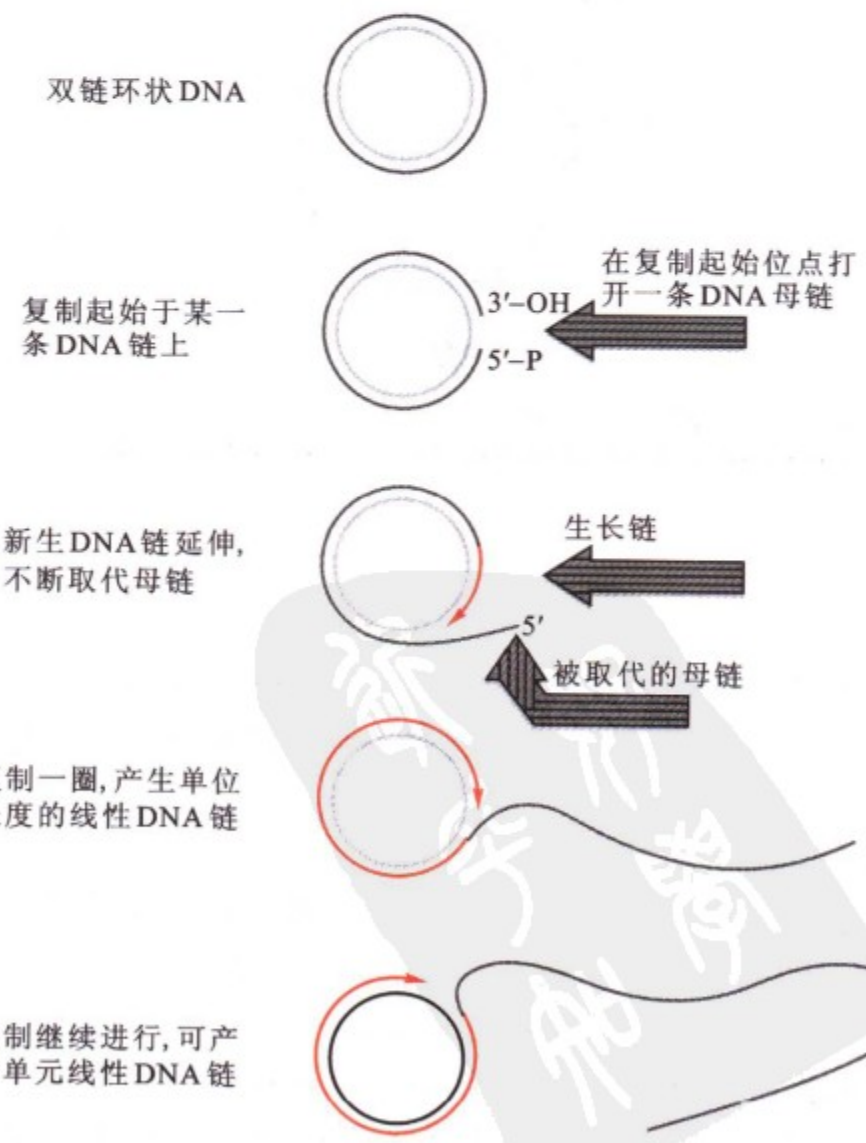



**图 2–21 环状 DNA 可以通过滚环式复制产生多单元单链 DNA**

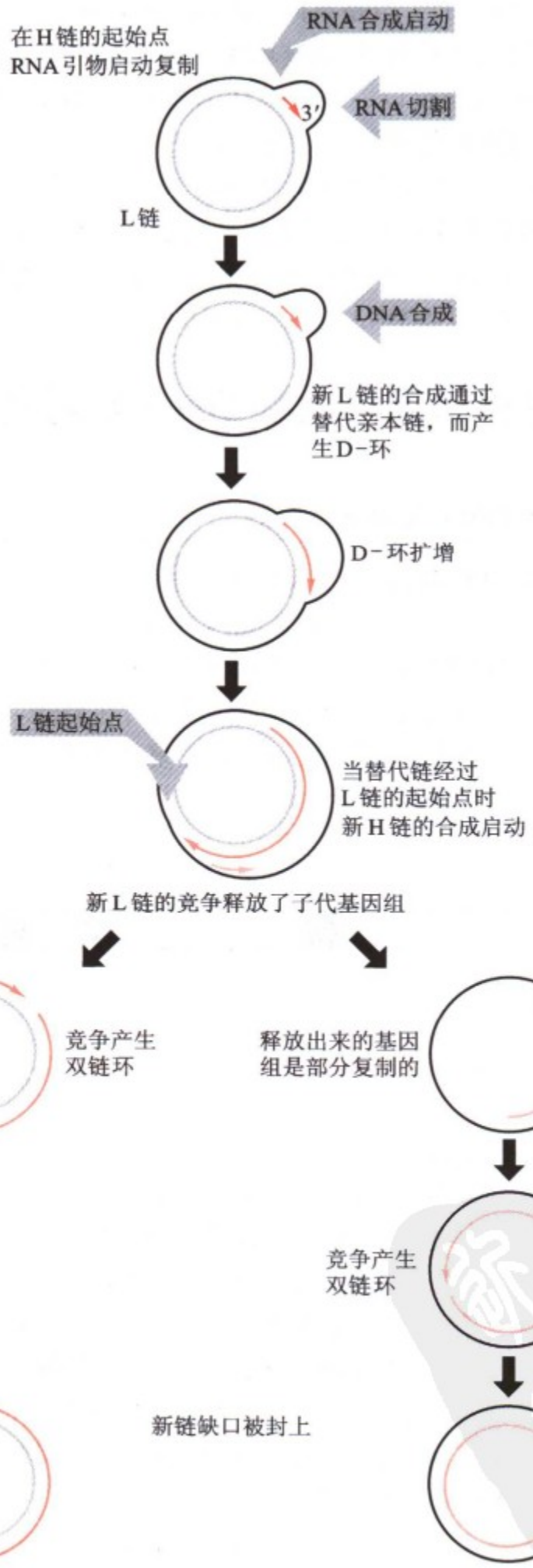



图 2–22 哺乳动物线粒体 DNA 复制从 D–环开始，每条链使用不同的起始点序列启动复制

(3) D-环型(D-loop) 也是单向复制的一种特殊方式。这种方式首先在动物线粒体 DNA 的复制中被发现,双链环在固定点解开进行复制。但两条链的合成是高度不对称的,最初仅以一条母链作为新链合成的模板,迅速合成出互补链,另一条链则成为游离的单链环(即 D-环)(图 2-22)。单个 D-环结构在哺乳动物线粒体中是一段开放的 500~600 碱基序列。维持 D-环结构的短链是不稳定的,它经常被解链和重新合成以维持该位点的双链开环结构。有些线粒体 DNA 拥有几个 D-环结构,反映出细胞有多个起始点。叶绿体 DNA 也采用同样的机制,高等植物的叶绿体 DNA 有两个 D-环结构。

## 2.4 原核生物和真核生物 DNA 复制的特点

### 2.4.1 原核生物 DNA 复制的特点

大肠杆菌基因组以双链环状 DNA 分子的形式存在,复制起始区位于其遗传图的 84 min 附近(图 2-23),复制起始点(OriC)含有 3 个 13 bp 的串联重复保守序列,GATCTNTTNTTTT,以及 4 个由 9 bp 的保守序列(TTATNCANA)组成的能够结合 DnaA 的起始结合位点。复制起始后,在 OriC 上形成的两个复制叉沿着整个基因组双向等速进行移动,其 DNA 复制的中间产物可形成一个 θ,两个复制叉在距起始点 180°处会合。

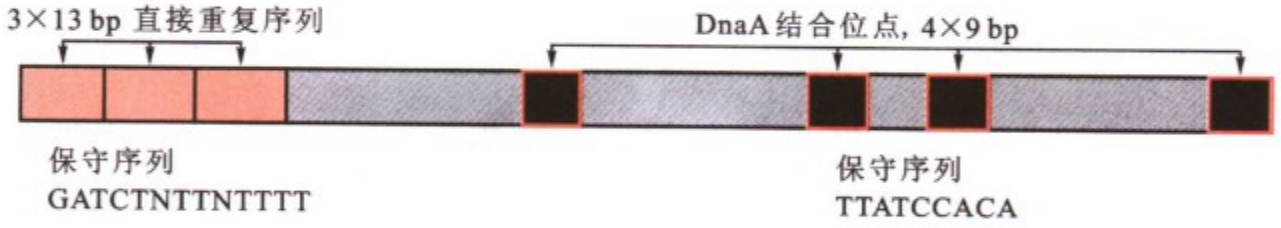



图 2-23 大肠杆菌 DNA 复制起始点(oriC)保守序列分布图

1. DNA 双螺旋的解旋

DNA 在复制时,其双链首先解开,形成复制叉,这是一个有多种蛋白质及酶参与的复杂过程。已经发现一些酶和蛋白质能使 DNA 双链变得易于解开,或者可以使超螺旋分子松弛。首先在拓扑异构酶 I 的作用下解开负超螺旋,并与解链酶共同作用,在复制起点处解开双链。参与解链的除一组解链酶外,还有 Dna 蛋白等,一旦局部解开双链,就必须有 SSB 蛋白来稳定解开的单链,以保证该局部结构不会恢复成双链,接着由引发酶等组成的引发体迅速作用于两条单链 DNA 上。不论是前导链还是后随链,都需要一段 RNA 引物以起始子链 DNA 的合成。现将其中重要的酶与蛋白质分述如下。

(1) DNA 解链酶 (DNA helicase) DNA 解链酶能通过水解 ATP 获得能量来解开双链 DNA。大部分 DNA 解链酶(包括大肠杆菌解链酶Ⅱ、Ⅲ、T4 噬菌体 dda、T4 基因 41 和人解链酶等) 可沿后随链模板的 5′→3′方向并随着复制叉的前进而移动,只有另一种解链(Rep 蛋白)是沿前导链模板的 3′→5′方向移动。

因此推测 Rep 蛋白和特定 DNA 解链酶分别在 DNA 的两条母链上协同作用,以解开双链 DNA。

(2) 单链结合蛋白 (single-strand binding protein,SSB 蛋白) 最初发现噬菌体 T4 的基因 32 蛋白可以在远低于解链温度时使双链 DNA 分开, 并牢牢地结合在单链 DNA 上, 后来发现许多生物中都有这类蛋白。从原核生物得到的 SSB 蛋白与 DNA 结合时还表现出协同效应:如第一个 SSB 蛋白结合到 DNA 上去的能力为 1,第二个 SSB 结合能力则可高达 $10^3$;真核生物牛、鼠和人细胞中的 SSB 蛋白与单链 DNA 结合时,则不表现上述协同效应。

SSB 蛋白的作用是保证被解链酶解开的单链在复制完成前能保持单链结构,它以四聚体形式存在于复制叉处,待单链复制完成后才离开,重新进入循环。所以,SSB 蛋白的作用是保持单链的存在,并没有解链的作用。

(3) DNA 拓扑异构酶(DNA topoisomerase) 大多数天然状态下的 DNA 分子都具有适度的负超螺旋,可以形成部分的单链结构,利于蛋白质与 DNA 的结合。在复制过程中,随着 DNA 的解旋,双螺旋的盘绕数 $T$ 减少,而超螺旋数 $W$ 增加,使正超螺旋增加,未解链部分的缠绕更加紧密,形成的压力使解链不能继续进行。拓扑异构酶能够消除解链造成的正超螺旋的堆积,消除阻碍解链继续进行的这种压力,使复制得以延伸。

2. DNA 复制的引发

所有 DNA 的复制都是从一个固定起始点开始的,而目前已知的 DNA 聚合酶都只能延长已存在的 DNA 链,而不能从头合成 DNA 链,那么一个新 DNA 的复制是怎样开始的呢?

研究发现, 所有的 DNA 聚合酶都从 3′羟基端起始 DNA 合成。DNA 复制时,往往先由引发酶(一种特殊的 RNA 聚合酶)在 DNA 模板上合成一段 RNA 链,它提供引发末端(被称为引物,primer),接着由 DNA 聚合酶从 RNA 引物 3′端开始合成新的 DNA 链。无论前导链还是后随链开始 DNA 合成时, 都需要 RNA 引物引发复制,只是对于前导链来说,这一引发过程比较简单,只要有一段 RNA 引物,DNA 聚合酶就能以此为起点一直合成下去。但对于后随链来说,引发过程就复杂多了,需要多种蛋白质和酶的协同作用,还牵涉冈崎片段的形成和连接。

后随链的引发过程往往由引发体(primosome)来完成。引发体由 6 种蛋白质 n,n′,n″,Dna B、C 和 I 共同组成,只有当引发前体(preprimosome)把这 6 种蛋白质合在一起并与引发酶(primase)进一步组装后形成引发体,才能发挥其功效。引发酶是 *dnaG* 基因的产物,是在特定环境下发挥作用的 RNA 聚合酶,仅用于合成 DNA 复制所需的一小段 RNA。引发体像火车头一样在后随链分叉的方向上前进, 并在模板上断断续续地引发生成后随链的引物 RNA 短链, 再由 DNA 聚合酶Ⅲ作用合成 DNA,直至遇到下一个引物或冈崎片段为止。由 RNase H 降解 RNA 引物并由 DNA 聚合酶Ⅰ将缺口补齐, 再由 DNA 连接酶将两个冈崎片段连在一起形成大分子 DNA。

(图 2-24) 以大肠杆菌为例解释了由 OriC 复制起始点处引发的 DNA 复制

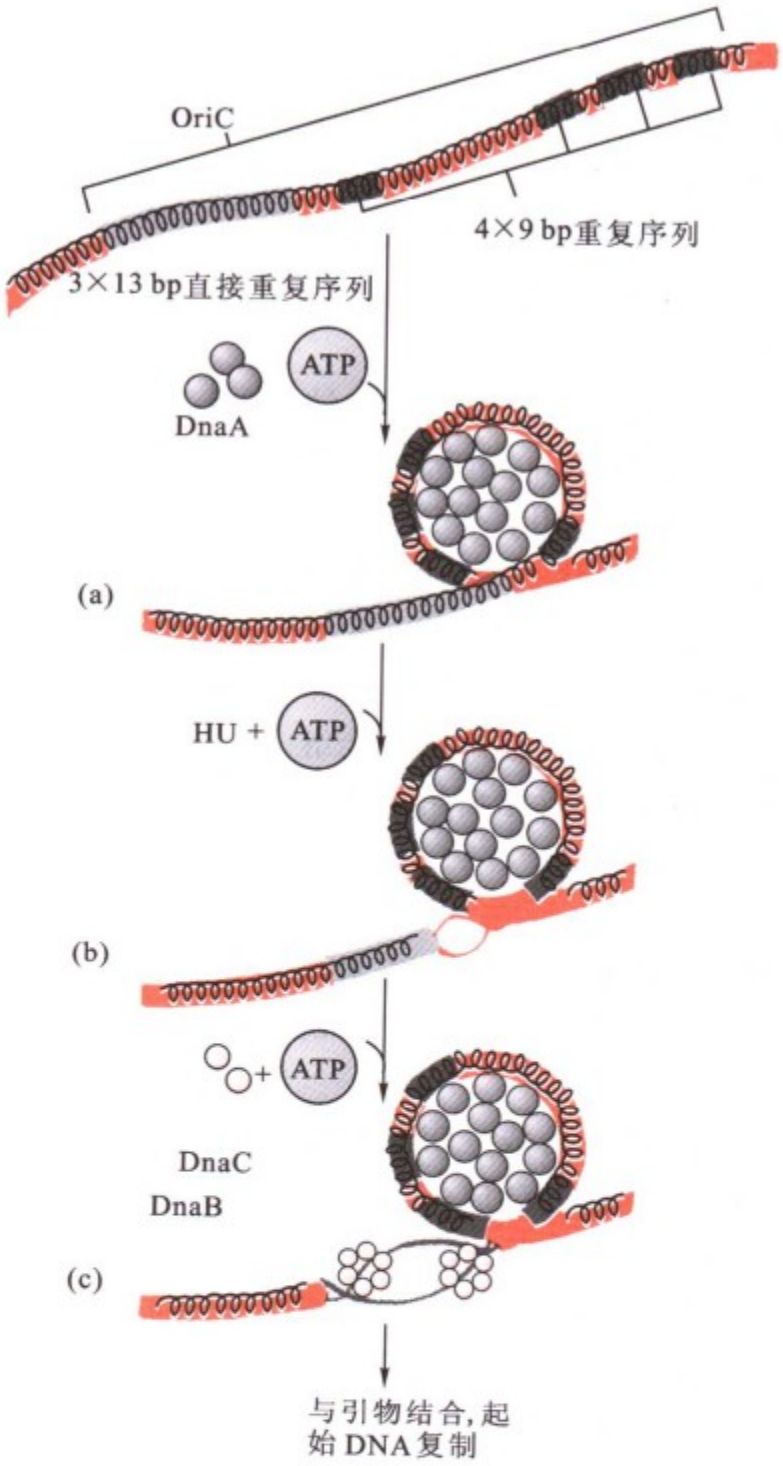



**图 2-24 由大肠杆菌 OriC 复制起始点处引发的 DNA 复制过程** (a)大约 20 个 DnaA 蛋白在 ATP 的作用下与 OriC 处的 4 个 9 bp 保守序列结合;(b)在 HU 蛋白和 ATP 的共同作用下,DnaA 复制起始复合物使 3×13 bp 直接重复序列变性,形成开链;(c)DnaB(解链酶)六体分别与单链 DNA 结合(需要 DnaC 的帮助),进一步解开 DNA 双链。

过程。在大肠杆菌中,一些蛋白质有序地作用于复制起始点。首先是在 ATP 的作用下,大约 20 个 DnaA 蛋白与 OriC 的 4 个 9 碱基保守序列结合,形成寡聚复合物;接着,在 DNA 结合蛋白 HU 蛋白和 ATP 的共同作用下,DnaA 复制起始复合物使 3 个 13 碱基重复序列变性,形成开链;在 DnaC 蛋白的协助下,六聚化的解链酶 DnaB 替换了 DnaA,与 DNA 链结合。目前认为 DnaB 可能具有两种构象,一种形式与双链 DNA 结合,另一种形式与单链 DNA 结合。这两种形式的转化可引发双链体的熔化,需要 ATP 的水解提供能量,即每解开一个碱基对需要水解一个 ATP。它在一个与双链区相连的单链区起始解链,使 DNA 双链进一步解开。

3. 冈崎片段与半不连续复制

由于 DNA 双螺旋的两条链是反向平行的,因此在复制叉附近解开的 DNA 链一条是 5′→3′方向,另一条是 3′→5′方向,两个模板极性不同。所有已知 DNA 聚合酶的合成方向都是 5′→3′,而不是 3′→5′,这就无法解释 DNA 的两条链是如何能够同时进行复制的。为了解释这一等速复制现象,日本学者冈崎(Okazaki)等提出了 DNA 的半不连续复制(semidiscontinuous replication)模型。他用 $^3H$ 脱氧胸苷短时间标记大肠杆菌,提取 DNA,变性后用超离心方法得到了许多 $^3H$ 标记的、被后人称为冈崎片段的 DNA,长度大约为 1 000~2 000 个碱基。现在已知一般原核生物的冈崎片段要长些,真核生物中的要短些。延长标记时间后,冈崎片段可转变为成熟 DNA 链,由此推断这些片段必然是复制过程的中间产物。此外,用 DNA 连接酶温度敏感突变株进行实验,在连接酶不起作用的温度下,可以观察到有大量小片段 DNA 的累积,说明 DNA 复制过程中至少有一条链首先合成较短的冈崎片段,然后再由连接酶连成大分子 DNA。

图 2-25 总结了两条新链不同的特征,根据它们不同性质分别将其称为前导链(leading strand)和后随链(lagging strand)。前导链 DNA 的合成以 5′→3′方向,随着亲本双链体的解开而连续进行复制;后随链在合成过程中,一段亲本

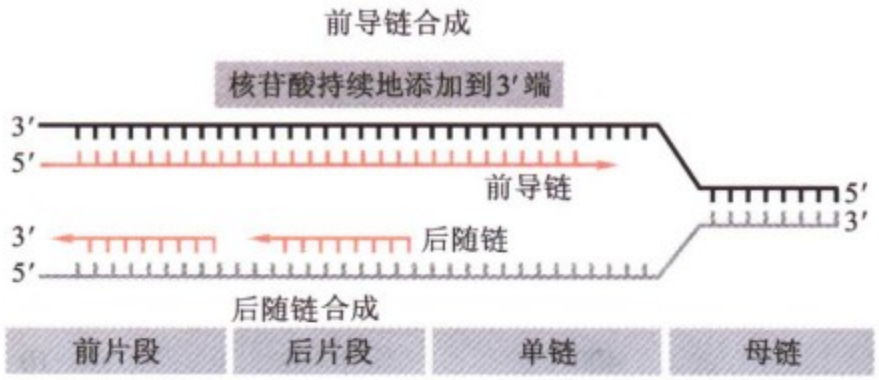



**图 2-25 DNA 合成的半不连续性** 前导链是持续合成的，而后随链则是以不连续的方式合成的。

DNA 单链首先暴露出来，然后以与复制叉移动相反的方向、按照 5′→3′方向合成一系列的冈崎片段，然后再把它们连接成完整的后随链。进一步研究还证明，这种前导链的连续复制和后随链的不连续复制在生物界是有普遍性的，因而称为双螺旋的半不连续复制。

4. 复制的终止

除 Tus 蛋白以外，链的终止看起来不需要太多蛋白质的参与。当复制叉前移，遇到约 22 个碱基的重复性终止子序列(Ter)时，Ter-Tus 复合物能使 DnaB 不再将 DNA 解链，阻挡复制叉的继续前移，等到相反方向的复制叉到达后停止复制，其间仍有 50~100 bp 未被复制，由修复方式填补空缺，而后两条链解开。在 DNA 拓扑异构酶Ⅳ的作用下使复制叉解体，释放子链 DNA(图 2-26)。

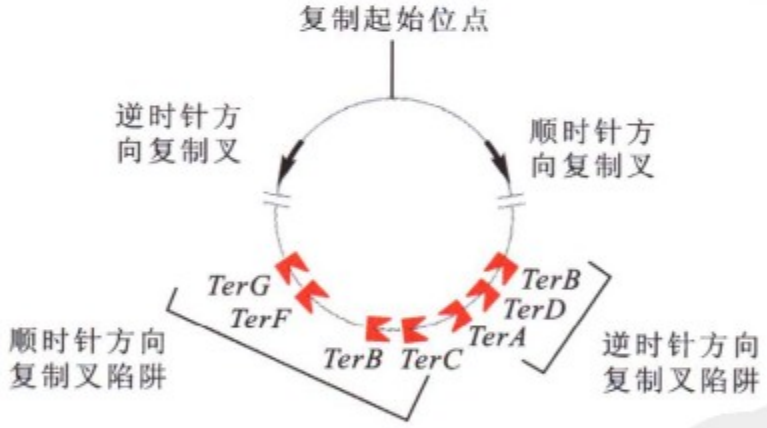



**图 2-26 大肠杆菌中 DNA 复制的终止**

5. DNA 聚合酶

现已发现在大肠杆菌中存在 DNA 聚合酶Ⅰ、Ⅱ、Ⅲ、Ⅳ和Ⅴ，有关 DNA 聚合酶Ⅰ、Ⅱ和Ⅲ(主要 DNA 聚合酶)的活性、功能及特点等归纳如表 2-11。DNA 聚合酶Ⅰ是第一个被鉴定出来的 DNA 聚合酶，但它不是复制大肠杆菌染色体的主要聚合酶。该蛋白可以被蛋白酶切成两个区域：占蛋白 2/3 的 C 端区域(相对分子质量 68 000)又称为 Klenow 片段，同时具有 DNA 聚合酶活性和 3′→5′核酸外切酶活性，既可合成 DNA 链，又能降解 DNA，保证了 DNA 复制的准确性。另外，它的 N 端区域(相对分子质量 35 000)具有 5′→3′核酸外切酶的活性，可作用于双链 DNA，又可水解 5′端或距 5′端几个核苷酸处的磷酸二酯键，因而

表 2-11 大肠杆菌 DNA 聚合酶Ⅰ、Ⅱ和Ⅲ的性质比较

| 性质 | 聚合酶Ⅰ | 聚合酶Ⅱ | 聚合酶Ⅲ |
|---|---|---|---|
| 3′→5′ | + | + | + |
| 5′→3′ | + | - | - |
| 新生链合成 | - | - | + |
| 相对分子质量( $\times 10^3$ ) | 103 | 90 | 900 |
| 细胞内分子数 | 400 | ? | 10 ~ 20 |
| 生物学活性 | 1 | 0.05 | 15 |
| 已知的结构基因 | Pol(A) | Pol(B) | Pol(C)(DnaE,N,Z,X,Q 等) |

该酶被认为在切除因紫外线照射而形成的嘧啶二聚体中起着重要的作用。它也可用以除去冈崎片段 5′端 RNA 引物,使冈崎片段间缺口消失,保证连接酶将片段连接起来。

DNA 聚合酶Ⅱ具有 5′→3′方向聚合酶活性,但酶活性很低。若以每分钟酶促核苷酸掺入 DNA 的转化率计算,只有 DNA 聚合酶Ⅰ的 5%,故也不是复制中主要的酶。其 3′→5′核酸外切酶活性可起校正作用。目前认为 DNA 聚合酶Ⅱ的生理功能主要是起修复 DNA 的作用。

DNA 聚合酶Ⅲ包含有 7 种不同的亚单位和 9 个亚基。生物活性形式为二聚体。它既有 5′→3′方向聚合酶活性,也有 3′→5′核酸外切酶活性。该酶的活性较强,为 DNA 聚合酶Ⅰ的 15 倍,DNA 聚合酶Ⅱ的 300 倍。它能在引物的 3′-OH 上以每分钟约 5 万个核苷酸的速率延长新生的 DNA 链,是大肠杆菌 DNA 复制中链延长反应的主导聚合酶。

DNA 聚合酶Ⅳ和Ⅴ分别由 *dinB* 和 *umuD′$_2$C* 基因编码,主要在 SOS 修复过程中发挥功能。

### 2.4.2 真核生物 DNA 复制的特点

真核生物 DNA 的复制与原核生物 DNA 复制有很多不同,例如,真核生物每条染色体上可以有多处复制起始点,而原核生物只有一个起始点;真核生物的染色体在全部完成复制之前,各个起始点上 DNA 的复制不能再开始,而在快速生长的原核生物中,复制起始点上可以连续开始新的 DNA 复制,表现为虽只有一个复制单元,但可有多个复制叉。

在真核生物细胞中,DNA 复制只是细胞周期的一部分,它只在 S 期进行。第一批复制子的激活标志着 S 期的开始。在之后的几个小时里,其余的复制子相继启动。真核生物的复制子相对较小,其长度为 40~100 kb。

真核生物中复制起始区有什么特点呢? 目前为止,酿酒酵母的复制起始位点已经被鉴定出来,长约 150 bp 左右,包括数个复制起始必需的保守区,将这个特殊序列作为一个部件来构建一个环状的 DNA 分子,这个核外 DNA 分子在酵母中是能够自主复制的。我们把酵母的复制起始位点称为自主复制序列(autonomously replicating sequences,ARS)。后来在其他的真核生物中也发现类

似于酵母 ARS 元件的顺序。不同 ARS 序列的共同特征是具有一个被称为 A 区的 11 个 A–T 碱基对的保守序列。真核生物 DNA 复制的起始需要起始点识别复合物（origin recognition complex，ORC）参与，人们首先在酿酒酵母中发现了 ORC，之后在裂殖酵母、果蝇和爪蟾中鉴定出相似的复合物，ORC 结合于 ARS，它是由 6 种蛋白质组成的启动复合物。

真核生物 DNA 复制叉的移动速度大约只有 50 bp/s，还不到大肠杆菌的 1/20，因此，人类 DNA 中每隔 30 000~300 000 个碱基就有一个复制起始位点。

已发现的真核生物 DNA 聚合酶有 15 种以上。在哺乳动物细胞中主要有 5 种 DNA 聚合酶，分别称为 DNA 聚合酶 α、β、γ、δ 和 ε，其特性如表 2-12。

表 2-12 真核生物 DNA 聚合酶的特性比较

| 性 质 | DNA 聚合酶 α | DNA 聚合酶 β | DNA 聚合酶 γ | DNA 聚合酶 δ | DNA 聚合酶 ε |
|---|---|---|---|---|---|
| 亚基数 | 4 | 1 | 2 | 2~3 | ≥1 |
| 在细胞内分布 | 核内 | 核内 | 线粒体 | 核内 | 核内 |
| 功能 | DNA 引物合成 | 损伤修复 | 线粒体 DNA 复制 | 主要 DNA 复制酶 | 复制修复 |
| 3′→5′外切 | 无 | 无 | 有 | 有 | 有 |
| 5′→3′外切 | 无 | 无 | 无 | 无 | 无 |

真核细胞的 DNA 聚合酶和细菌 DNA 聚合酶基本性质相同，均以 dNTP 为底物，需 $Mg^{2+}$ 激活，聚合时必须有模板链和具有 3′–OH 末端的引物链，链的延伸方向为 5′→3′。但真核细胞的 DNA 聚合酶一般都不具有核酸外切酶活性，推测一定有另外的酶在 DNA 复制中起校对作用。

DNA 聚合酶 α 的功能主要是引物合成，即能起始前导链和后随链的合成。它是以复合体的形式存在的，因其具有引发、延伸链的双重功能，所以又被称为 pol α 的引发酶。DNA 聚合酶 β 活性水平稳定，可能主要在 DNA 损伤的修复中起作用，属于高忠实性修复酶。DNA 聚合酶 δ 是主要负责 DNA 复制的酶，参与前导链和后随链的合成。而 DNA 聚合酶 ε 与后随链合成有关，在 DNA 合成过程中核苷切除以及碱基的切除修复中起着重要的作用，而且，它在细胞的重组过程中可能具有某些功能。DNA 聚合酶 γ 在线粒体 DNA 的复制中发挥作用。

除了上述 5 种主要的 DNA 聚合酶外，真核生物中还存在 ζ、η、ι 和 κ 等几种 DNA 聚合酶，它们承担着修复损伤的功能，但这些修复酶的忠实性都很低。

有关真核生物 DNA 复制比较公认的说法是：在复制叉上存在一个 DNA 聚合酶 α/引发酶复合体和另外两个 DNA 聚合酶复合体。这两个聚合酶中一个是 DNA 聚合酶 δ，另一个是第二个 DNA 聚合酶 δ 或 DNA 聚合酶 ε，前者延伸前导链，后者合成后随链的冈崎片段。冈崎片段 RNA 引物的去除分为两步。首先，一种可特异性切除 DNA–RNA 杂合底物的 RNA 酶 H1 发挥核酸内切酶

的活性,在靠近 RNA 与 DNA 的连接处切开引物,由具备 5′-3′核酸外切酶活性的 FEN1 蛋白降解 RNA 片段。最后,由 DNA 连接酶 I 将相邻的冈崎片段连接起来。

真核生物染色体复制时,前导链可连续复制直到模板链的末端,释放出完整的子代染色单体;而后随链以不连续的方式复制,必须有特殊的机制来安排 5′端部分 DNA 链的复制问题,以免该链子代 DNA 的 5′端序列逐步缩短。真核生物能够通过形成端粒结构和具有反转录酶活性的端粒酶来防止 DNA 复制时后随链缩短而产生的染色体部分缺失。端粒酶能利用自身携带的 RNA 链作为模板,以 dNTP 为原料,以反转录的方式催化合成模板后随链 5′端 DNA 片段或外加重复单位(如人染色体端粒为 TTAGGG),以维持端粒一定的长度,从而防止染色体的缺短损伤。

### 2.4.3 DNA 复制的调控

原核细胞的生长和增殖速度取决于培养条件,但在生长增殖速度不同的细胞中 DNA 链延伸的速度几乎是恒定的,只是复制叉的数量不同。迅速分裂的细胞具较多复制叉,而分裂缓慢的细胞复制叉较少并出现复制的间隙。细胞内复制叉的多少决定了复制起始频率的高低,这可能是原核细胞复制的调控机制。复制起始频率的直接调控因子是蛋白质和 RNA。

1. 大肠杆菌染色体 DNA 的复制调控

染色体的复制与细胞分裂一般是同步的,但复制与细胞分裂不直接偶联。复制起始不依赖于细胞分裂,而复制的终止则能引发细胞分裂。在一定生长速度范围内,细胞与染色体的质量之比相对恒定,这是由活化物、阻遏物和去阻遏物及它们的相互作用所制约的。复制的功能单位,即复制子与转录的操纵子相似,由起始物位点和复制起点两部分组成。起始物位点编码复制调节蛋白质,复制起点与调节蛋白质相互作用并启动复制。起始物位点突变使复制停止并导致细胞死亡。这些条件致死突变包括 dnaA、dnaC 和 dnaT 突变株,证明上述基因编码的调节蛋白质通过与复制复合物的相互作用确定复制起始频率和复制方式。

DNA 复制的调控主要发生在起始阶段,一旦开始复制,如没有意外的阻力,就一直可以复制下去直到完成。大肠杆菌的复制起点有 OriC 和 OriH 两种,其中 OriC 是首选的复制起点,而 OriH 是在 RNaseH 缺失突变株中发现的一系列复制起点。这两种起点起始复制的机制和调控方式不同。

2. ColE1 质粒 DNA 的复制调控

ColE1 是一个 6 646 bp 的小质粒,在宿主细胞内拷贝数为 20~30。ColE1 DNA 复制不依赖于其本身编码的蛋白质,而完全依靠宿主 DNA 聚合酶 I。质粒 DNA 编码两个负调控因子 Rop 蛋白和反义 RNA($RNA_1$),它们控制了起始 DNA 复制所必需的引物合成。引物 RNA 前体的转录起始于复制起点上游 555 个核苷酸处,需经 RNaseH 加工后产生有 555 个核苷酸的引物,然后由 DNA 聚合酶 I 在引物的 3′端起始 DNA 合成。$RNA_1$ 的编码区在引物 RNA 编码区的 5′端,转

录方向与引物 RNA 相反,因此与引物 RNA 的 5′端互补(图 2-27)。$RNA_1$ 通过氢键配对与引物 RNA 前体相互作用,阻止了 RNaseH 加工引物前体,使其不能转化为有活性的引物而对复制起负调控作用。

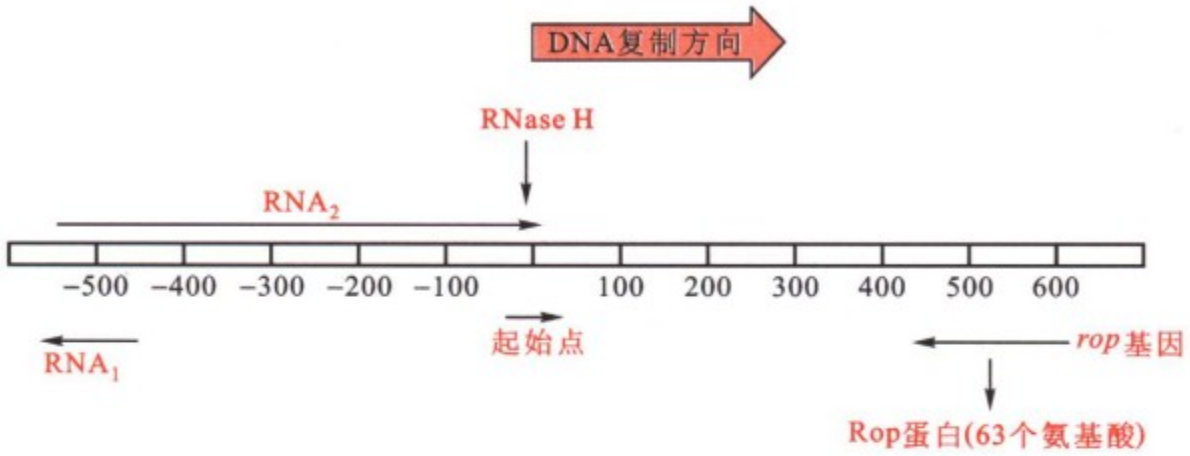



图 2-27 ColE1 质粒 DNA 复制起始部位调控因子之间关系简图

$RNA_1$ 不仅控制质粒的拷贝数而且决定了质粒的不相容性。$RNA_1$ 与引物 RNA 分子的相互作用是可逆的,因此细胞内 $RNA_1$ 的浓度决定了 ColE1 质粒的复制起始频率。另一个负调控因子 Rop 蛋白能提高 $RNA_1$ 与引物前体的相互作用,从而加强了 $RNA_1$ 的负调控作用。

3. 真核细胞 DNA 的复制调控

真核细胞的生活周期可分为 4 个时期:G1、S、G2 和 M 期。G1 是复制预备期,S 为复制期,G2 为有丝分裂准备期,M 为有丝分裂期。DNA 复制只发生在 S 期。真核细胞中 DNA 复制有 3 个水平的调控:

(1) 细胞生活周期水平调控　也称为限制点调控,即决定细胞停留在 G1 期还是进入 S 期。许多外部因素和细胞因子参与限制点调控。促细胞分裂剂、致癌剂、外科切除等都可诱发细胞由 G1 期进入 S 期。一些细胞质因子如四磷酸二腺苷和聚 ADP-核糖也可诱导 DNA 的复制。

(2) 染色体水平调控　决定不同染色体或同一染色体不同部位的复制子按一定顺序在 S 期起始复制,这种有序复制的机制还不清楚。

(3) 复制子水平调控　决定复制的起始与否。这种调控从单细胞生物到高等生物是高度保守的。此外,真核生物复制起始还包括转录活化、复制起始复合物的合成和引物合成等阶段,许多参与复制起始蛋白的功能与原核生物中相类似。酵母染色体复制只发生于 S 期,各个复制子按专一的时间顺序活化,在 S 期的不同阶段起始复制。研究由 3 个复制起点(ARS1、ARS2 和 10Z)构建的质粒发现,酵母复制起始受时序调控,也受 α 因子和 *cdc* 基因调控。已克隆的 ARS 片段中都有 14 bp 的核心序列,其中序列为 A(或 T)TT-TATPuTTA(或 T)的 11 个核苷酸是高度保守的,这个区域的点突变使 ARS 失去复制起始功能。

## 2.5 DNA 的修复

由于染色体 DNA 在生命过程中占有至高无上的地位,DNA 复制的准确性

表 2-13 大肠杆菌中 DNA 的修复系统

| DNA 修复系统 | 功 能 |
|---|---|
| 错配修复 | 恢复错配 |
| 切除修复(碱基、核苷酸) | 切除突变的碱基和核苷酸片段 |
| 重组修复 | 复制后的修复,重新启动停滞的复制叉 |
| DNA 直接修复 | 修复嘧啶二体或甲基化 DNA |
| SOS 系统 | DNA 的修复,导致变异 |

以及 DNA 日常保养中的损伤修复就有着特别重要的意义。表 2-13 是存在于大肠杆菌中的 DNA 修复系统。

## 2.5.1 错配修复

一旦在 DNA 复制过程中发生错配，细胞能够通过准确的错配修复(mismatch repair)系统识别新合成链中的错配并加以校正,DNA 子链中的错配几乎完全能被修正,充分反映了母链序列的重要性。因此,错配修复系统对 DNA 复制忠实性有很大的贡献。该系统识别母链的依据来自 Dam 甲基化酶，它能使位于 5′-GATC 序列中腺苷酸的 $N^6$ 位甲基化。一旦复制叉通过复制起始位点，母链就会在开始 DNA 合成前的几秒钟至几分钟内被甲基化。此后,只要两条 DNA 链上碱基配对出现错误,错配修复系统就会根据“保存母链,修正子链”的原则,找出错误碱基所在的 DNA 链，并在对应于母链甲基化腺苷酸上游鸟苷酸的 5′位置切开子链（图 2-28），再根据错配碱基相对于 DNA 切口的方位启动如图 2-29 中所示两条修复途径之一，合成新的子链 DNA 片段。

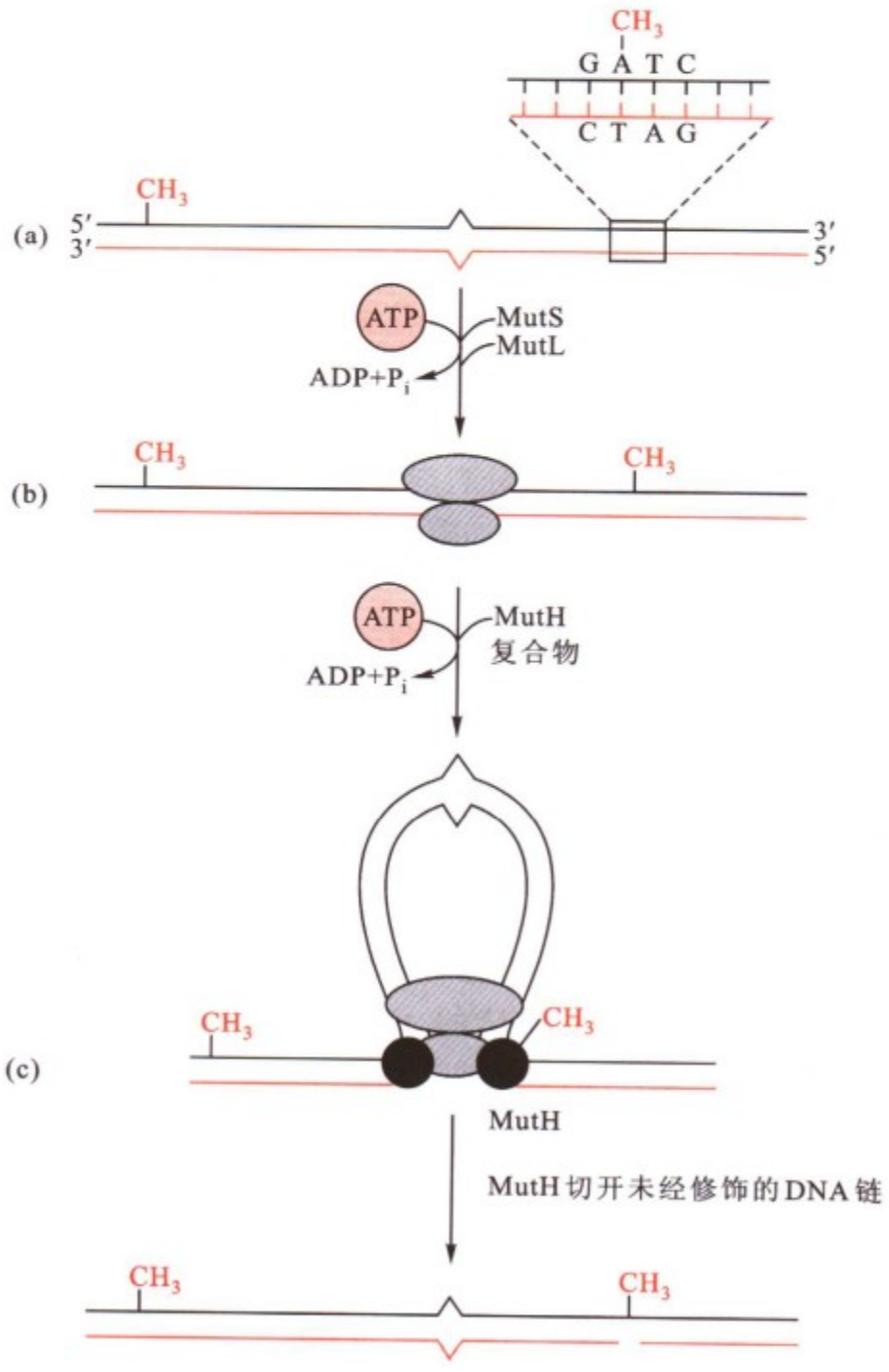



**图 2-28 根据母链甲基化原则找出错配碱基过程图示** (a)发现碱基错配;(b)在水解 ATP 的作用下,MutS,MutL 与碱基错配位点的 DNA 双链结合;(c)Muts-MutL 在 DNA 双链上移动，发现甲基化 DNA 后由 MutH 切开非甲基化的子链。

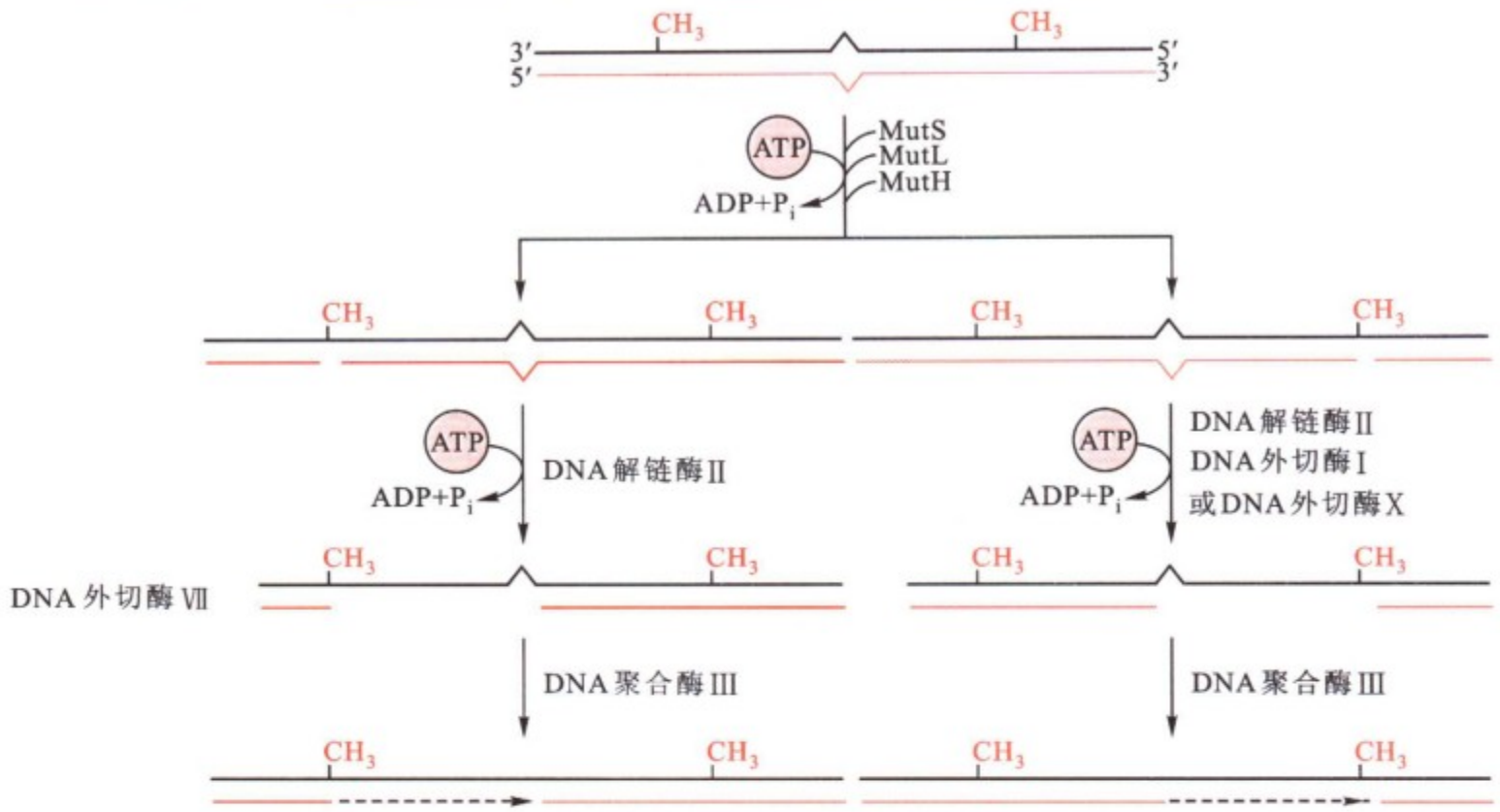



**图 2-29 碱基错配修复过程示意图** 当错配碱基位于切口 3′下游端时，在MutL-MutS、解链酶 II 、DNA 外切酶 VII或 RecJ 核酸酶的作用下，从错配碱基 3′下游端开始切除单链 DNA 直到原切口，并在 Pol III和 SSB 的作用下合成新的子链片段。若错配碱基位于切口的 5′上游端，则在 DNA 外切酶 I 或 X 的作用下，从错配碱基 5′上游端开始切除单链 DNA 直到原切口，再合成新的子链片段。

## 2.5.2 切除修复

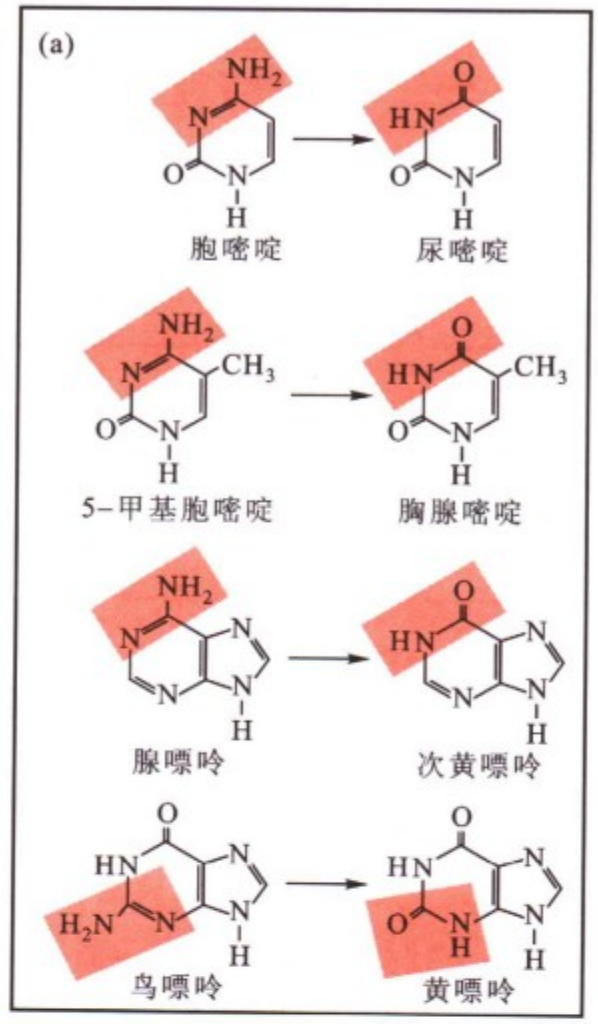
切除修复(excision repair)，主要分为以下两种。

1. 碱基切除修复(base-excision repair)

研究发现，所有细胞中都带有不同类型、能识别受损核酸位点的糖苷水解酶，它能特异性切除受损核苷酸上的 *N*-*β*-糖苷键，在 DNA 链上形成去嘌呤或去嘧啶位点，统称为 AP 位点(图 2-30)。一类 DNA 糖苷水解酶一般只对应于某一特定类型的损伤，如尿嘧啶糖苷水解酶就特异性识别 DNA 中胞嘧啶自发脱氨形成的尿嘧啶，而不会水解 RNA 分子中尿嘧啶上的 *N*-*β*-糖苷键。DNA 分子中一旦产生了 AP 位点，AP 核酸内切酶就会把受损核苷酸的糖苷-磷酸键切开，并移去包括 AP 位点核苷酸在内的小片段 DNA，由 DNA 聚合酶 I 合成新的片段，最终由 DNA 连接酶把两者连成新的被修复的 DNA 链。

2. 核苷酸切除修复(nucleotide-excision repair)

当 DNA 链上相应位置的核苷酸发生损伤，导致双链之间无法形成氢键，则由核苷酸切除修复系统负责修复。图 2-31 分别是大肠杆菌（左）和人类细胞(右)中的修复过程。损伤发生后，首先由 DNA 切割酶(excinuclease)在已损伤的核苷酸 5′ 和 3′ 位分别切开磷酸糖苷键，产生一个由 12~13 个核苷酸(原核生物)或 27~29 个核苷酸(人类或其他高等真核生物)组成的小片段，移去小片段



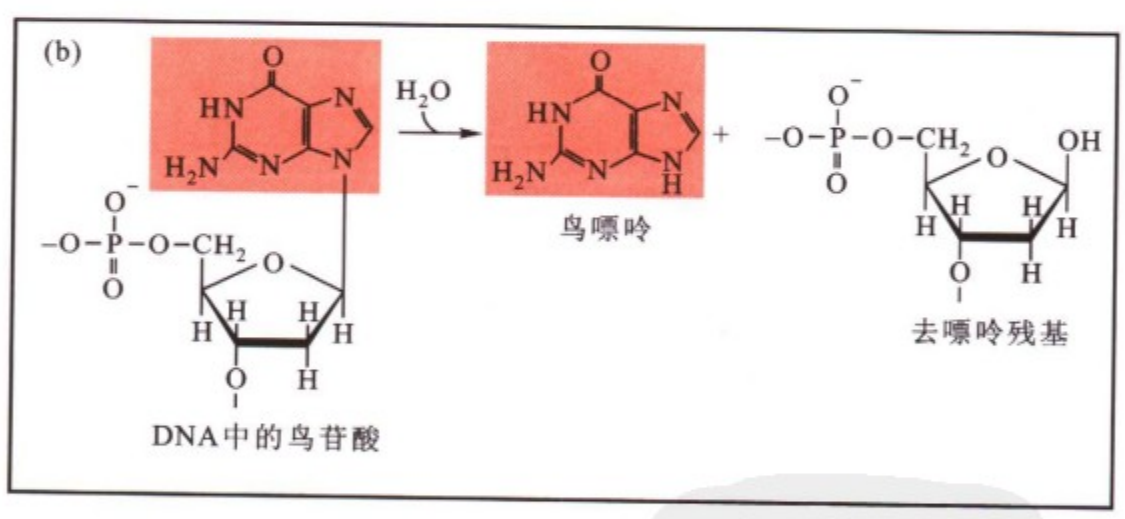




**图 2–30 DNA 分子中常见的几种核苷酸非酶促转变反应** (a)脱氨基反应；(b)脱嘌呤反应(*N*–*β*–糖苷键被水解)。

后由 DNA 聚合酶 I (原核)或 ε(真核)合成新的片段，并由 DNA 连接酶完成修复中的最后一道工序。

### 2.5.3 重组修复

重组修复(recombinant repair)又被称为“复制后修复”，它发生在复制之后。机体细胞对在复制起始时尚未修复的 DNA 损伤部位可以先复制再修复，即先跳过该损伤部位，在新合成链中留下一个对应于损伤序列的缺口。该缺口由 DNA 重组来修复：先从同源 DNA 母链上将相应核苷酸序列片段移至子链缺口

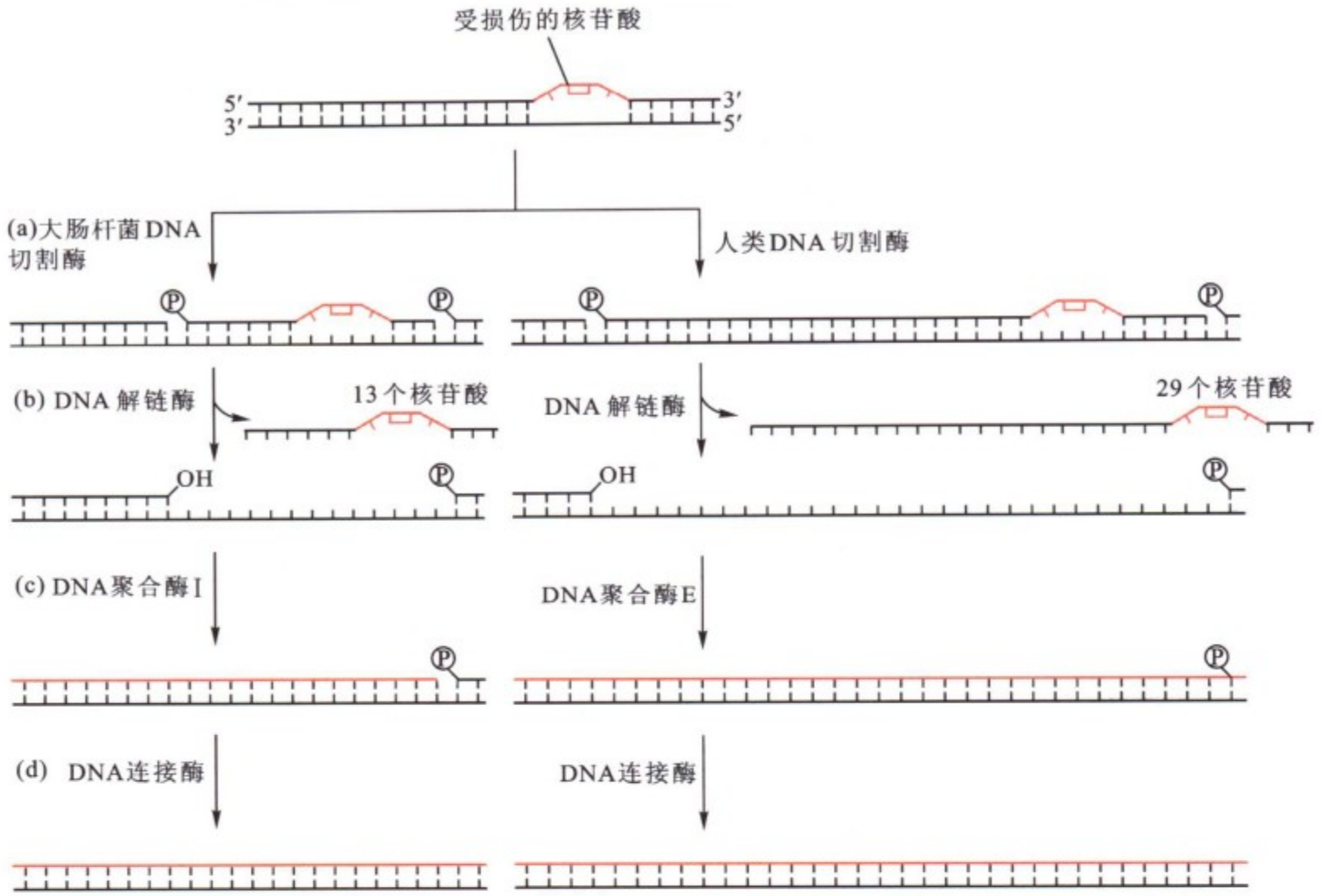



**图 2-31 大肠杆菌(左)和人类细胞(右)中的核苷酸切除修复过程示意图** 在原核生物中受损伤核苷酸 3′端的第 5 位,5′端的第 8 位磷酸糖苷键分别被 DNA 切割酶切开。在人类细胞中,受损伤核苷酸 3′端第 6 位,5′端的第 22 位磷酸糖苷键分别被 DNA 切割酶切开。

处,然后再用新合成的序列补上母链空缺。大肠杆菌的 *rec* 基因编码主要的重组修复系统,它的一个主要作用是重新启动停滞的复制叉。

### 2.5.4 DNA 的直接修复

生物体内还存在多种 DNA 损伤以后直接修复(direct repair)而并不需要切除碱基或核苷酸的机制。直接修复是把损伤的碱基回复到原来状态的一种修复。最普遍的例子是在 DNA 光解酶(photolyase)的作用下把在光下或经紫外线照射形成的环丁烷胸腺嘧啶二体及 6-4 光化物(6-4photoproduct)还原成为单体的过程(图 2-32)。生物体内还广泛存在着使 $O^6$-甲基鸟嘌呤脱甲基化的甲基转移酶(图 2-33),以防止形成 G-T 配对。

### 2.5.5 SOS 反应

SOS 反应(SOS response)是细胞 DNA 受到损伤或复制系统受到抑制的紧急情况下,细胞为求生存而产生的一种应急措施。SOS 反应包括诱导 DNA 损伤修复、诱变效应、细胞分裂的抑制以及溶原性细菌释放噬菌体等,细胞癌变也与 SOS 反应有关。

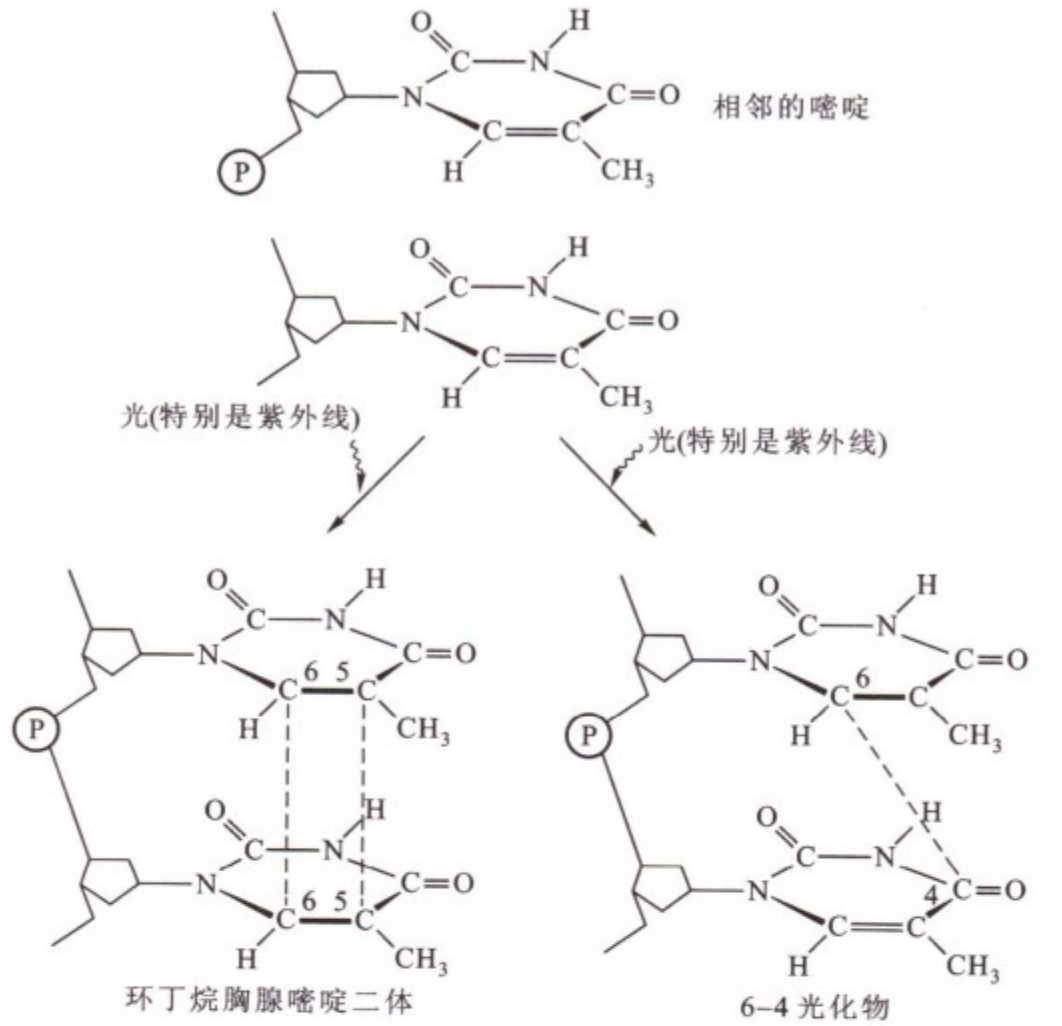



图 2-32 紫外光诱发形成嘧啶二体

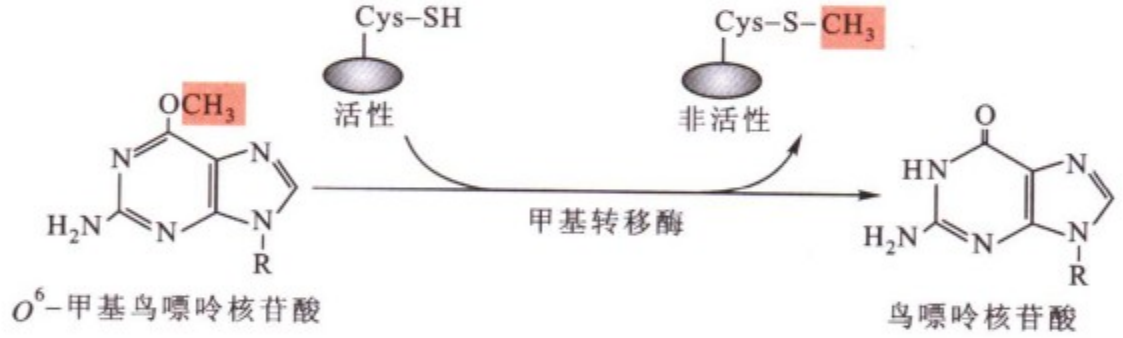



图 2-33 $O^6$-甲基转移酶将 $O^6$-甲基化鸟嘌呤恢复成鸟嘌呤

SOS 反应广泛存在于原核和真核生物中，主要包括两个方面：①DNA 的修复；②产生变异。前者利于细胞的存活，具有重要意义，而后者可能产生不利的后果，如导致细胞的癌变。

## 2.6 DNA 的转座

DNA 的转座，或称移位(transposition)，是由可移位因子(transposable element)介导的遗传物质重排现象。与 DNA 的同源重组相比，转座现象发生的频率虽然要低得多，但它仍然有着十分重要的生物学意义，这不仅因为它能说明在细菌中发现的许多基因缺失或倒转现象，而且它常被用于构建新的突变体。已经发现“转座”这一命名并不十分准确，因为在转座过程中，可移位因子的一

个拷贝常常留在原来位置上,在新位点上出现的仅仅是它的拷贝。因此,转座有别于同源重组,它依赖于 DNA 的复制。

## 2.6.1 转座子的分类和结构特征

转座子(transposon,Tn)是存在于染色体 DNA 上可自主复制和移位的基本单位。Tn 最初是由 Barbara McClintock 于 20 世纪 40 年代在玉米的遗传学研究中发现的。几十年的研究结果表明转座子存在于所有的生物体内,人类基因组中有约 35%以上的序列为转座子序列,其中大部分与疾病相关。

转座子分为两大类:插入序列(insertional sequence,IS)和复合型转座子。

1. 插入序列(IS 因子)

插入序列是最简单的转座子,它不含有任何宿主基因。它们是细菌染色体或质粒 DNA 的正常组成部分。一个细菌细胞常带有少于 10 个 IS 序列。转座子常常被定位到特定的基因中,造成该基因突变。科学上用标准命名法对这些突变进行编号,如 λ:IS1 表示有一个 IS1 序列插入到 λ 噬菌体基因组内。IS 序列都是可以独立存在的单元,带有介导自身移动的蛋白,也可作为其他转座子的组成部分。

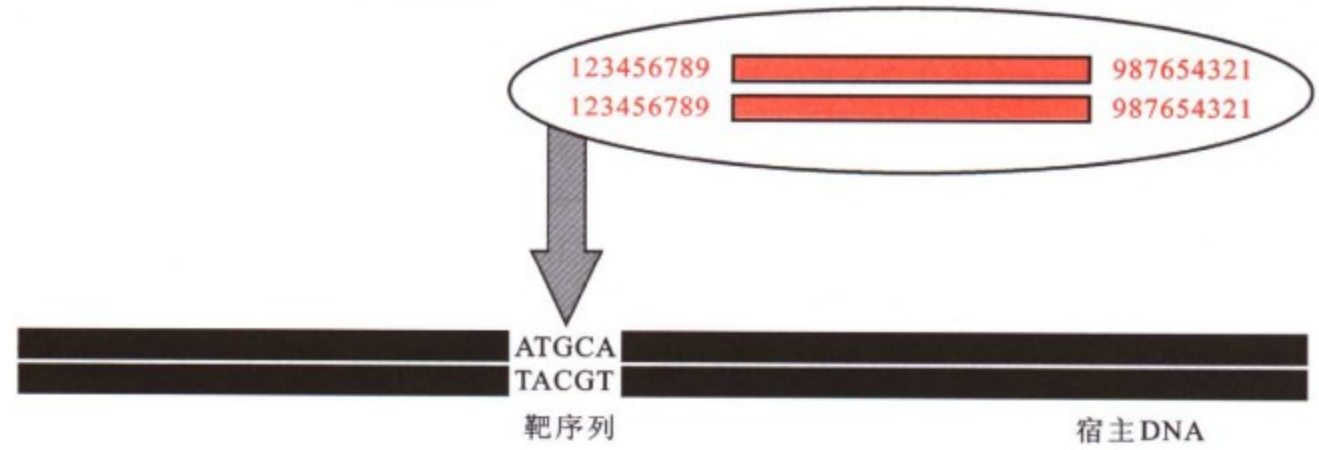
表 2-14 列出了部分常见的 IS 序列及其特征,它们都是很小的 DNA 片段(约 1 kb),末端具有倒置重复序列,转座时往往复制宿主靶位点一小段(4~15 bp)DNA,形成位于 IS 序列两端的正向重复区(图 2-34)。除 IS1 以外,所有已知 IS 序列都只有一个可译框架,翻译起始位点紧挨着第一个倒置重复区,终止点位于第二个倒置重复区或重复区附近。IS1 含有两个分开的读码框,只有移码通读才能产生功能型转座酶。一般情况下,每个 IS 转座频率是 $10^{-4}$~$10^{-3}$/世代,恢复频率则低得多,约为 $10^{-10}$~$10^{-6}$/世代。

表 2-14 IS 序列的结构特征比较

| | 长度(bp) | 两端倒置重复区(bp) | 靶位点正向重复区(bp) | 靶位点 |
|---|---|---|---|---|
| IS1 | 768 | 23 | 9 | 随机 |
| IS2 | 1 327 | 41 | 5 | 有热点 |
| IS4 | 1 428 | 18 | 11 或 12 | AAAN20TTT |
| IS5 | 1 195 | 16 | 4 | 有热点 |
| IS10R | 1 329 | 22 | 9 | NGCTNAGCN |
| IS50R | 1 531 | 9 | 9 | 有热点 |
| IS903 | 1 057 | 18 | 9 | 未知 |

2. 复合型转座子

复合型转座子(composite transposon)是一类带有某些抗药性基因(或其他宿主基因)的转座子,其两翼往往是两个相同或高度同源的 IS 序列,表明 IS 序列插入到某个功能基因两端时就可能产生复合转座子(图 2-35)。一旦形成复合转座子,IS 序列就不能再单独移动,因为它们的功能被修饰了,只能作为复合



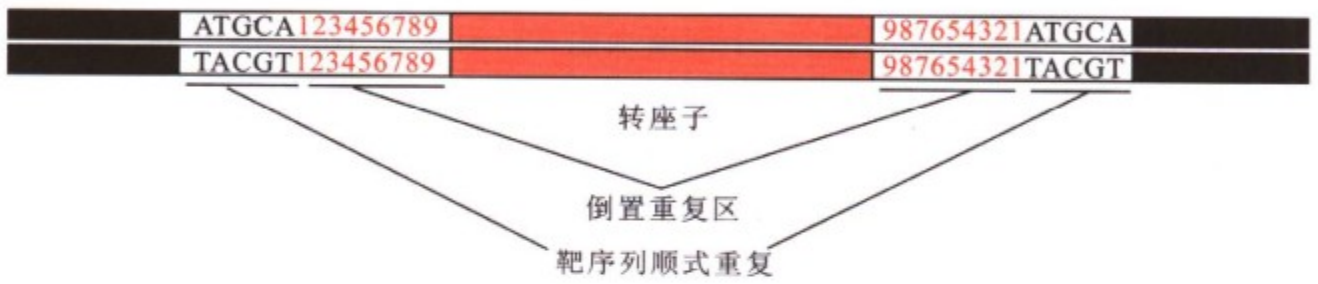




图 2-34 DNA 转座的一般模式

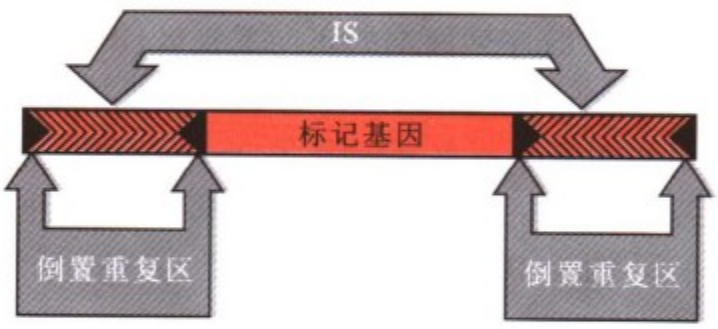



图 2-35 复合转座子的两端往往各有一个 IS 序列，在每个 IS 序列两侧各有倒置重复序列

体移动。大部分情况下，这些转座子的转座能力是由 IS 序列决定和调节的。

除了末端带有 IS 序列的复合转座子以外，还存在一些没有 IS 序列的、体积庞大的转座子(5 000 bp 以上)——TnA 家族。这类转座子带有 3 个基因，其中一个编码 $\beta$-内酰胺酶(AmpR)，另两个则是转座作用所必需的。图 2-36 是转座子 TnA 结构模式。所有 TnA 类转座子两翼都带有 38 碱基的倒置重复序列。

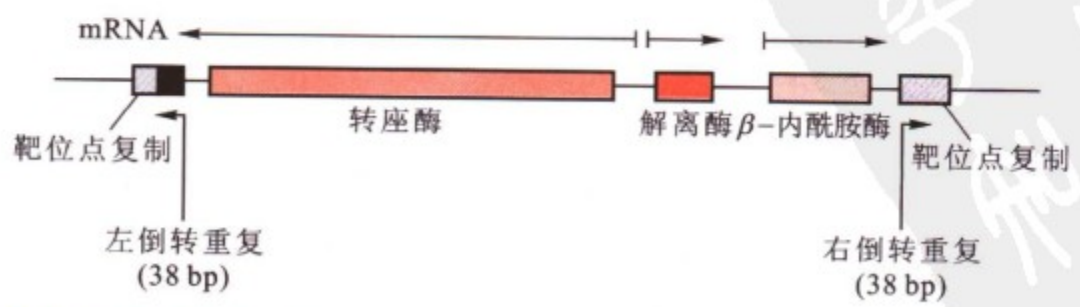



图 2-36 转座子 TnA 的结构示意图

## 2.6.2 真核生物中的转座子

转座子不仅存在于原核细胞(如大肠杆菌)和低等真核细胞(如酵母)中,也同样存在于高等真核生物中。如玉米和果蝇中就发现了多个在基因组内随机分布而且能重复移动的转座子。大量研究证实,几乎所有高等生物基因组中都存在类似转座因子序列,高等真核生物基因组的流动性可能比原核生物还要大些。

1. 玉米中的控制因子(controlling element)

早在 20 世纪初,遗传学家就已发现玉米中存在决定体细胞变异的控制因子(事实上就是转座子)。半个多世纪前 McClintock 在美国康奈尔大学和冷泉港实验室所进行的开创性工作,打破了孟德尔关于基因固定排列于染色体上的概念,尽管在当时几乎不能被科学界所接受,30 多年后却终于在分子水平上得到证实,并使她获得了诺贝尔生理医学奖。

可以把玉米细胞内的控制因子归纳为两大类:一类是自主性因子(autonomous element);另一类是非自主性因子(nonautonomous element)。前者具有自主剪接和转座的功能,后者单独存在时是稳定的,不能转座,当基因组中存在与非自主性因子同家族的自主性因子时,它才具备转座功能,成为与自主性因子相同的转座子。同一家族的自主性因子能为非自主性因子的转座提供反式作用蛋白(转座酶),而不同家族间无此反应。研究发现,玉米转座子同样具有典型的 IS 特征——在转座子的两翼有两个倒转重复序列,在靶 DNA 插入位点有两个短的正向重复序列。

下面主要介绍在玉米中已研究得较清楚的两个控制因子:

(1) Ac-Ds 系统(activator-dissociation system) 即激活-解离系统,它通过非复制型机制进行转座。在这个体系中,自主转座子 Ac 长 4 563 bp,转录生成单一 RNA,其产物为转座酶。成熟 mRNA 长 3 500 bp,并含有一个长 807 个密码子的可译框架,被 4 个内含子分隔成 5 个外显子。Ac 转座子的两翼有 11 bp 的反向重复序列,在其靶 DNA 位点复制形成 8 bp 正向重复。Ac 能够自主转座,并形成不稳定的基因突变,但不使染色体断裂;它能使 Ds 因子活化、转座,并通过 Ds 控制结构基因的表达。Ac 的作用还表现为剂量效应,Ac 的增加可以推迟 Ds 因子的解离和转座。

Ds 元件属于非自主因子,又称解离因子,与 Ac 属于同一家族的控制因子。已知所有 Ds 都是 Ac 转座子的缺失突变体,如 Ds9 只缺失了 194 bp,而 Ds6 缺了约 2.5 kb,因而失去转座酶功能。当 Ac 因子存在时,能活化 Ds,使其在基因组内转座或插入结构基因之内,导致基因失活或改变结构基因的表达水平,也可使染色体特定部位断裂,引起缺失或重组。

非自主性转座子虽然缺失内源序列,但其两端转座特征序列却是完整的,只要细胞内有相应的转座酶活性,它就能恢复转座功能。与上述 Ds 不同的是 Ds1 家族,它们只与 Ac 两翼的 11 bp 倒转重复区同源,暗示不同的转座酶只识别转座子两端的倒转重复区,或者包括一小部分内源序列。

（2）Spm-dSpm 系统(suppressor-promoter-mutator) 即抑制-促进-增变系统。Spm 是自主性因子,又称增强因子。Spm 和 En 只有不到 10 个碱基的差异,两端各有一个 13 碱基的倒置重复序列,在其靶 DNA 位点复制形成 3 碱基正向重复。Spm 和 En 的转录区(即 *tnpA* 基因)有 8 300 bp,包括 11 个外显子,成熟 mRNA 为 2 500 个碱基,编码了含 621 个氨基酸、能与转座子近末端区 12 bp 多拷贝保守序列相结合的蛋白质。研究发现,*tnpA* 基因产物的主要功能是切割转座子序列。另外两个可译框架位于 *tnpA* 基因的第一个内含子中。因为编码这两个蛋白的 6 000 碱基长的 mRNA 产生于转录产物的剪接过程中,该 mRNA 在细胞中的含量大约只有 *tnpA* 基因产物的 1%。由 *tnpB* 基因编码的这两个蛋白可能直接与转座子两端 13 碱基倒转重复区相结合，引起 DNA 切割和转移。Spm 具有转座、整合和解离活性,dSpm(defective Spm)是非自主元件,所有 dSpm 都是功能型 Spm 的缺失突变体(图 2-37)。有意思的是,如果缺失突变发生在与内含子 1 相对应的区域,破坏了 *tnp*B 基因,该突变体的转座功能并不完全消失,仅产生弱 Spm,说明 *tnp*B 基因产物的功能至少可以被部分替代。Spm-dSpm 在功能上与 Ac-Ds 系统相似,也可引入基因的插入突变,影响结构基因表达,还能导致染色体断裂。

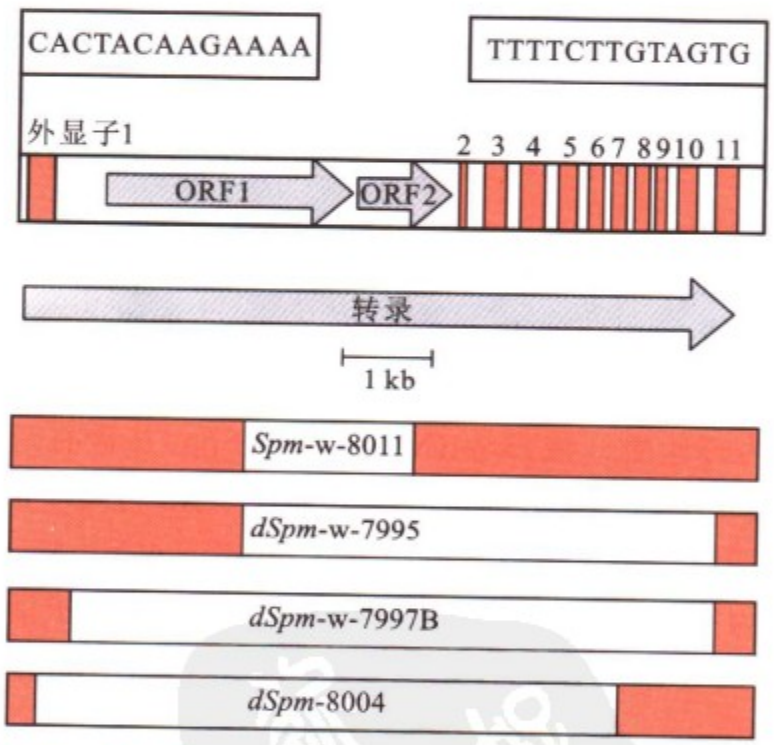



图 2-37 转座子 Spm/En 和 dSpm 非自主转座子的结构比较

2. 果蝇中的转座子

果蝇中的 Copia 转座子具有数百碱基的末端正向重复和类似于酵母 Ty 转座子及 RNA 肿瘤前病毒的结构(图 2-38)。Copia 类的主要特征是它们能以高速度转录并产生有多聚腺苷化末端的全长 RNA,序列分析表明,它只具有一个长的可译框架,其基因产物与 RNA 肿瘤病毒反转录酶有较高的同源性。

最有意思的果蝇转座子可能要算诱发杂种不育(hybrid dysgenesis)的 P 转座子(P element)了。实验证明,果蝇中几乎所有的杂种不育都是由于 P 转座子

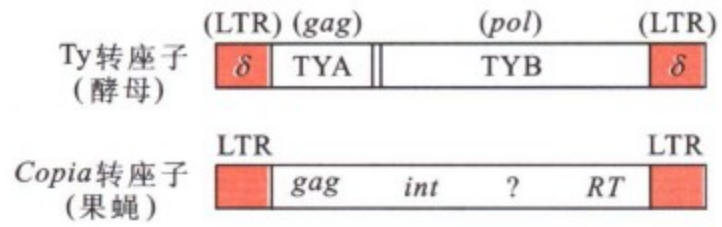



**图 2–38 真核生物细胞内存在不少结构上类似于反转录病毒的反转录转座子** 这类转座子都编码一个与反转录病毒中的反转录酶高度同源的转座酶，左右两边都是与病毒边界序列 LTR 同源的重复序列。

插入基因组 W 位点而引起的。所有 P 转座子两翼都有 31 碱基倒置重复序列，P 转座子转座导致靶 DNA 复制产生 8 碱基正向重复序列。图 2–39 是常见功能型 P 转座子示意图，其原初转录产物为约 2.5 或 3.0 kb，包括外显子 0、1、2、3 及 3 个内含子。体细胞中，只有前两个内含子能被顺利切除，产生包括外显子 0~2 的功能型 mRNA，并被翻译成一个相对分子质量 $6.6\times10^4$ 的转座阻遏物。在卵细胞中，内含子 3 能被切除，所产生的成熟 mRNA 包括全部 4 个外显子并被翻译成 $8.7\times10^4$ 转座酶，导致 P 转座子转座和配子体败育。

有两个方面的证据表明第三个内含子的切除是发生转座所必需的。第一，如果突变内含子 3 的末端序列，该 P 转座子就丧失转座活性；第二，如果从核基因中切除与内含子 3 相对应的 DNA 序列，以保证外显子 3 成为功能 mRNA 的有效组分，该 P 转座子就能同时在体细胞和卵细胞内实施转座。那么，为什么内含子 3 能在卵细胞内被切除却在体细胞内被保留呢？原来体细胞中存在一个与外显子 3 特异性结合的蛋白，这种蛋白与 RNA 的结合妨碍了内含子 3 的剪接。

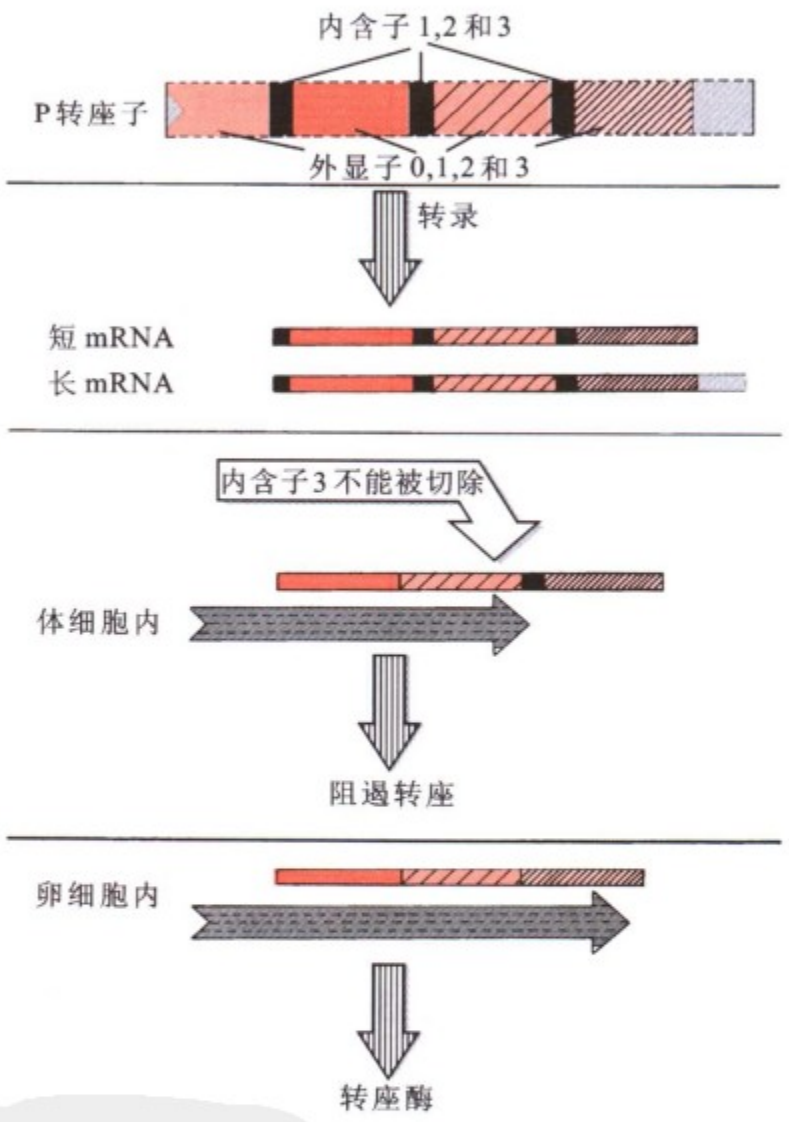



**图 2–39 果蝇转座子对杂种后代不育性的影响分析**

P 因子的转座属于非复制型，能引起原 P 位点断裂。可以把果蝇分为 P(父本贡献型)和 M(母本贡献型)两大类，只有雄性 P 型果蝇和雌性 M 型果蝇杂交的后代才表现为杂种不育(图 2–40)。科学家发现，果蝇的细胞质因子(cytotype)决定了杂种后代的育性。在雄性 P 果蝇×雌性 P 果蝇、雄性 P 果蝇×雌性 M 果蝇和雄性 M 果蝇×雌性 P 果蝇 3 组杂种后代中，只有第二种杂交导致后代不育，因为雌性 M 果蝇细胞内缺失一个 $6.6\times10^4$ 的转座阻遏物，引起雄性来源染色体上 P 转座子自由转移，并造成卵细胞败育。在其余两组杂交中，由母本提供的正常细胞质因子都能抑制 P 转座子的转座，保证后代可育。

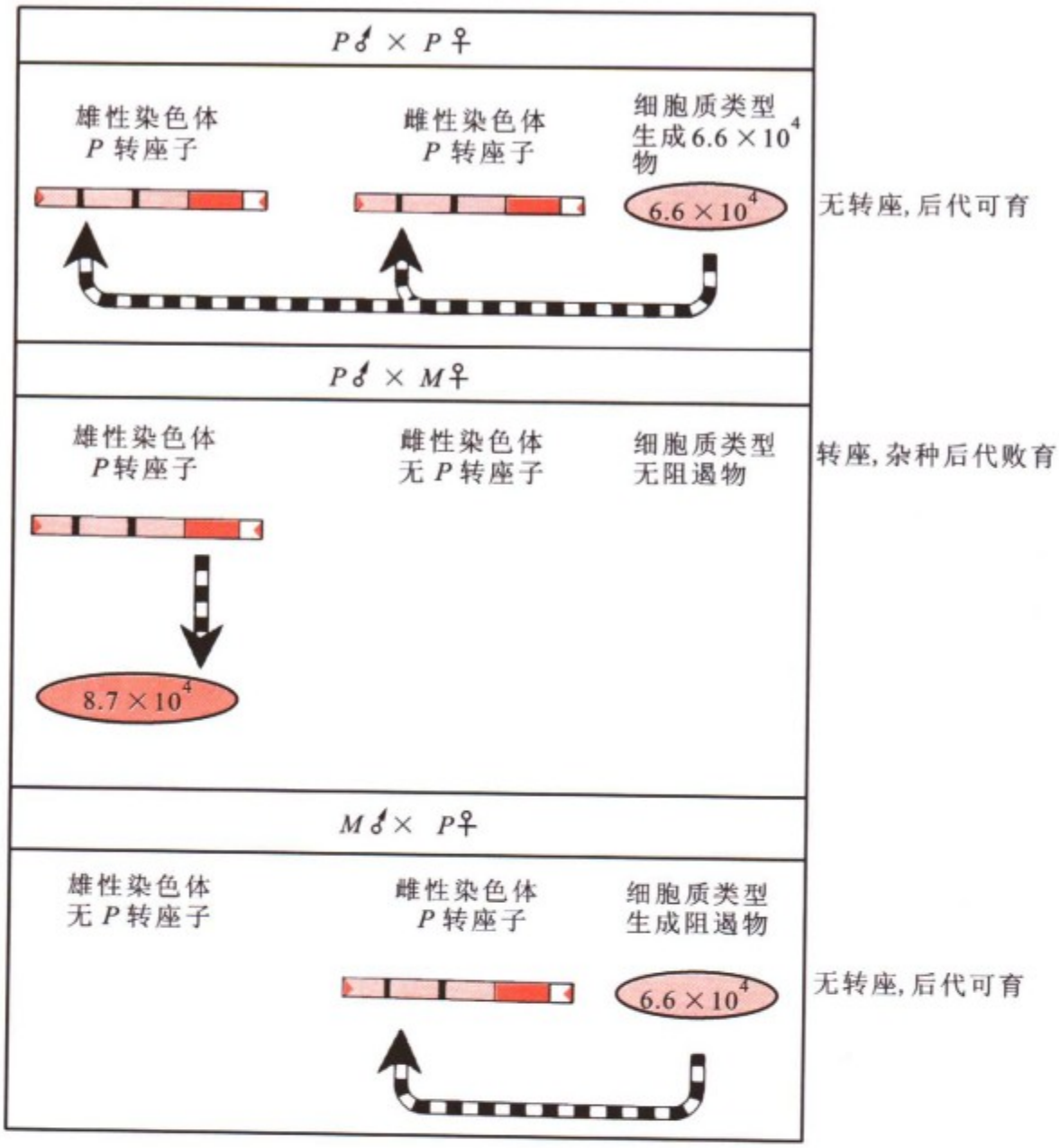



图 2-40 P转座子与杂种不育

## 2.6.3 转座作用的机制

转座时发生的插入作用有一个普遍的特征，那就是受体分子中有一段很短的(3~12 bp)、被称为靶序列的 DNA 会被复制，使插入的转座子位于两个重复的靶序列之间。不同转座子的靶序列长度不同，但对于一个特定的转座子来说，它所复制的靶序列长度都是一样的，如 IS1 两翼总有 9 bp 的靶序列，而 Tn3 两端总有 5 bp 的靶序列(图 2-41)。研究发现，靶序列的复制可能起源于特定内切酶所造成的 DNA 黏性末端。

转座可被分为复制型和非复制型两大类。顾名思义，在复制型转座中，整个转座子被复制了，所移动和转位的仅仅是原转座子的拷贝。转座酶(transposase)和解离酶(resolvase)分别作用于原始转座子和复制转座子。TnA 类转座主要是这种形式。与此相对应，在非复制型转座中，原始转座子作为一个可移动的实体直接被移位，IS 序列、Mu 及 Tn5 等都以这种方式进行转座。

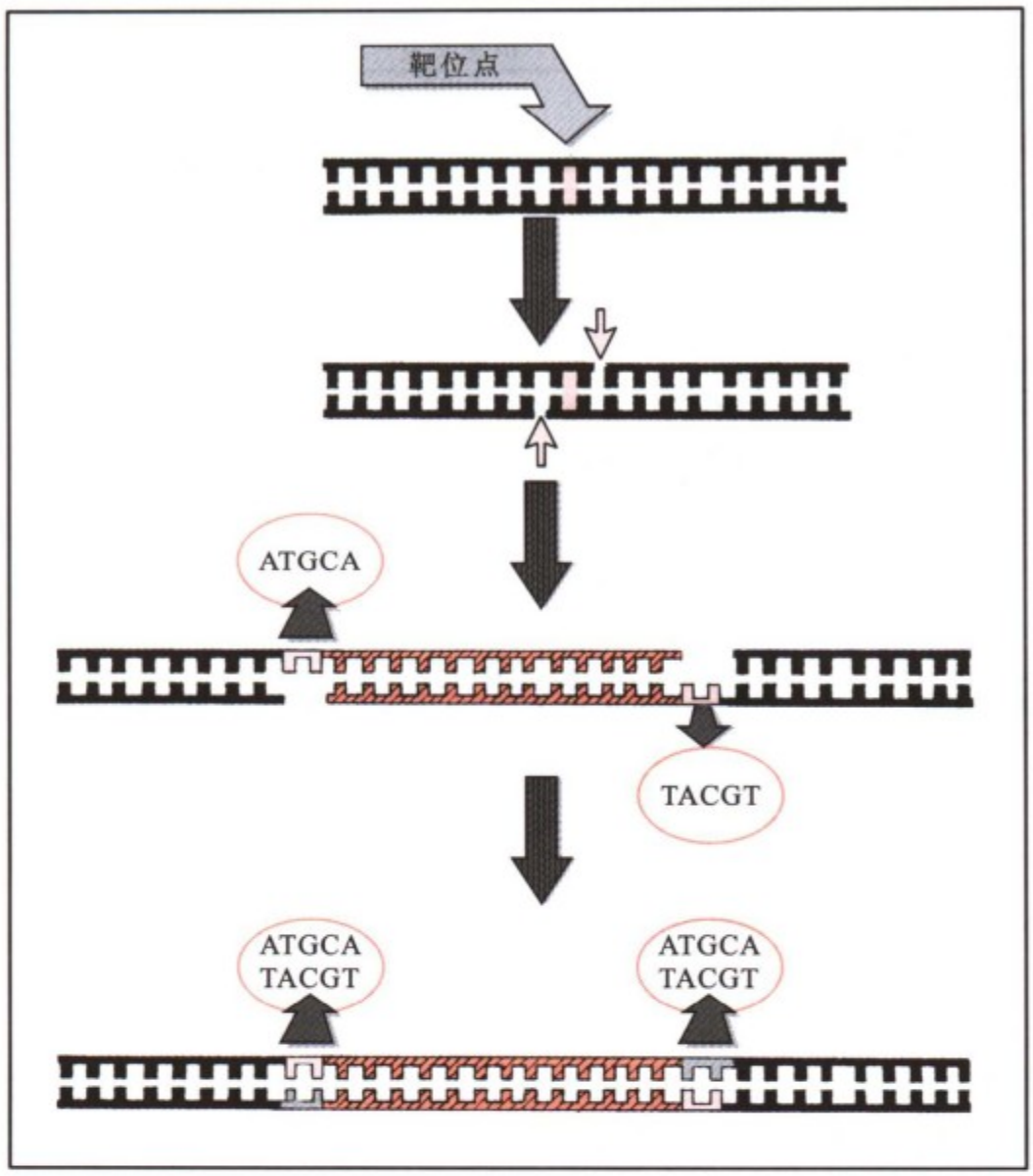



图 2-41 靶位点直接重复序列的形成

## 2.6.4 转座作用的遗传学效应

转座子引发了许多遗传变异，如基因重排及质粒-染色体 DNA 整合等，DNA 转座的遗传学效应主要有以下几方面：

1. 转座引起插入突变

各种 IS、Tn 转座子都可以引起插入突变。如果插入位于某操纵子的前半部分，就可能造成极性突变，导致该操纵子后半部分结构基因表达失活。

2. 转座产生新的基因

如果转座子上带有抗药性基因，它一方面造成靶 DNA 序列上的插入突变，同时也使这个位点产生抗药性。

3. 转座产生的染色体畸变

当复制性转座发生在宿主 DNA 原有位点附近时，往往导致转座子两个拷贝之间的同源重组，引起 DNA 缺失或倒位。若同源重组发生在两个正向重复转座区之间，就导致宿主染色体 DNA 缺失；若重组发生在两个反向重复转座区之间，则引起染色体 DNA 倒位（图 2-42）。

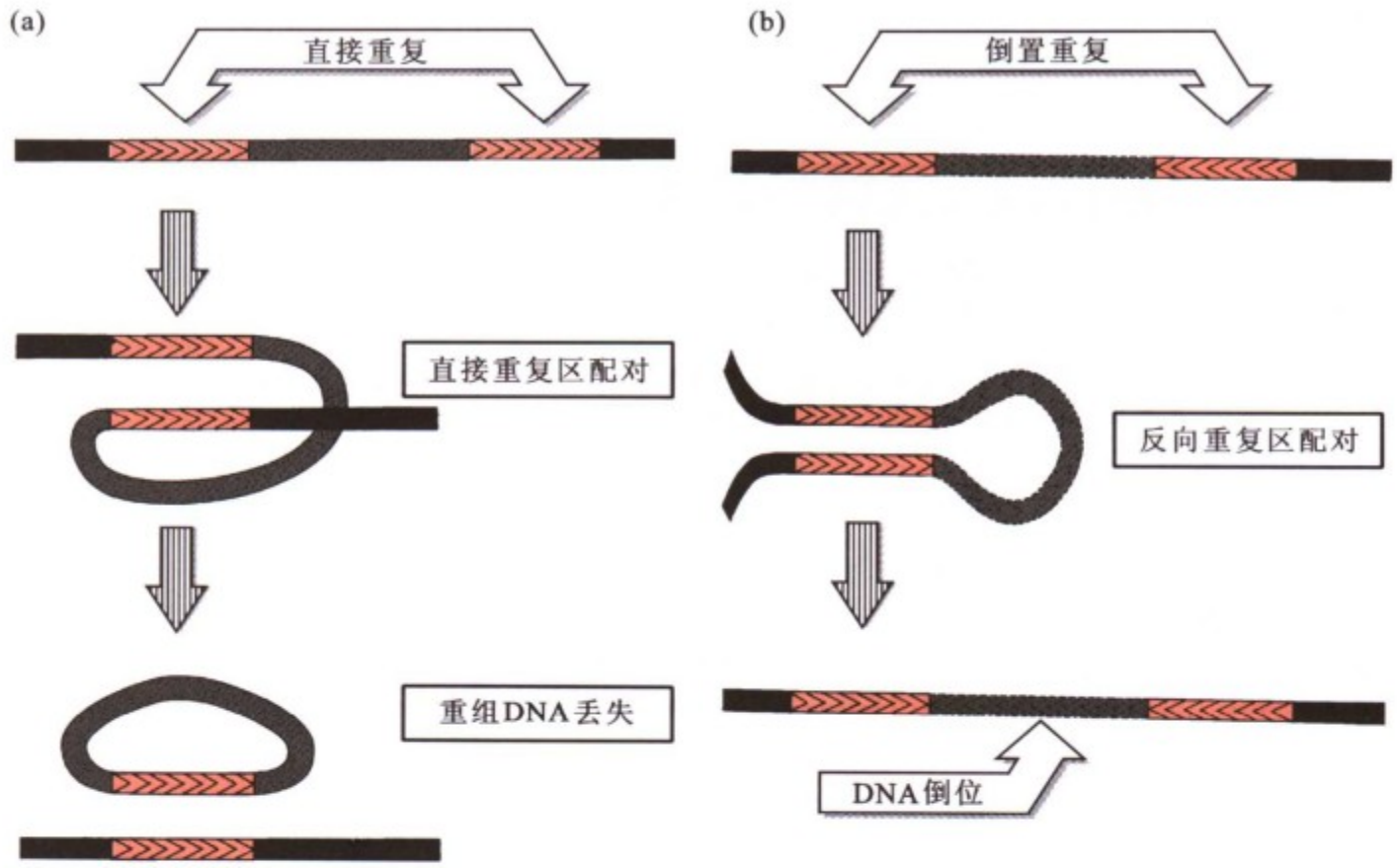



图 2-42 由转座子引起的染色体 DNA 缺失(a)或倒位(b)

4. 转座引起的生物进化

由于转座作用，使原来在染色体上相距甚远的基因组合到一起，构建成一个操纵子或表达单元，可能产生有新的生物学功能的基因和新的蛋白质分子。

## 思考题

1. 染色体具备哪些作为遗传物质的特征?
2. 什么是核小体? 简述其形成过程。
3. 简述真核生物染色体的组成及组装过程。
4. 简述 DNA 的一、二、三级结构特征。
5. 原核生物 DNA 具有哪些不同于真核生物 DNA 的特征?
6. 简述 DNA 双螺旋结构及其在现代分子生物学发展中的意义。
7. DNA 复制通常采取哪些方式?
8. 简述原核生物 DNA 的复制特点。
9. 真核生物 DNA 的复制在哪些水平上受到调控?
10. 细胞通过哪几种修复系统对 DNA 损伤进行修复?
11. 什么是转座子? 可分为哪些种类?
12. 请说说插入序列与复合型转座子之间的异同。

# 主要参考文献


1. Arents G, Burlingame R W, Wang B C, et al. The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. Proc. Nat. Acad. Sci. U.S.A .1991, 88: 10148-10152.
2. Calos M P, Miller J H. Transposable elements. Cell, 1980, 20: 579~595.
3. Drake J W, Charlesworth B, Charlesworth D, Crow J F. Rates of spontaneous mutation. Genetics, 1998, 148: 1667-1686.
4. Fedoroff N. Transposons and genome evolution in plants. Proc. Nat. Acad. Sci. U.S.A. 2000, 97: 7002-7007.
5. Holmes F. Meselson, Stahl, and the Replication of DNA: A History of the Most Beautiful Experiment in Biology. Yale University Press, 2001.
6. Kornberg R D. Chromatin structure: a repeating unit of histones and DNA. Science, 1974, 184: 868-871.
7. Luger K, Mader A W, Richmond R K. et al. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature, 1997, 389: 251-260.
8. Marki H. Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. Ann. Rev. Genet. 2002, 36: 279~303
9. McGhee J D, Felsenfeld G. Nucleosome structure. Ann. Rev. Biochem. 1980, 49: 1115-1156.
10. Meselson M, Stahl F W. The replication of DNA in Escherichia coli. Proc. Nat. Acad. Sci. U.S.A .1958, 44: 671-682.

# 3 生物信息的传递(上)
## ——从 DNA 到 RNA


- 3.1 RNA 转录的基本过程
- 3.2 转录机器的主要成分
- 3.3 启动子与转录起始
- 3.4 原核与真核生物 mRNA 的特征比较
- 3.5 终止和抗终止
- 3.6 内含子的剪接、编辑、再编码及化学修饰


现代分子生物学的最基本原理是:基因作为唯一能够自主复制、永久存在的单位,其生物功能是以蛋白质的形式表达出来的。所以我们说,DNA 序列是遗传信息的贮存者,它通过自主复制得到永存,并通过转录生成信使 RNA,翻译生成蛋白质的过程来控制生命现象。

基因表达包括转录(transcription)和翻译(translation)两个阶段,转录是指拷贝出一条与 DNA 链序列完全相同(除了 T→U 之外)的 RNA 单链的过程,是基因表达的核心步骤;翻译是指以新生的 mRNA 为模板,把核苷酸三联遗传密码子翻译成氨基酸序列、合成多肽链的过程,是基因表达的最终目的。因为只有 mRNA 所携带的遗传信息才被用来指导蛋白质生物合成,所以人们一般用 U、C、A、G 这 4 种核苷酸而不是 T、C、A、G 的组合来表示遗传性状。

DNA 是贮藏遗传信息的最重要的生物大分子。DNA 分子中的核苷酸排列顺序不但决定了胞内所有 RNA 及蛋白质的基本结构,还通过蛋白质(酶)的功能间接控制了细胞内全部有效成分的生产、运转和功能发挥。我们把与mRNA 序列相同的那条 DNA 链称为编码链(coding strand)或称有意义链(sense strand),并把另一条根据碱基互补原则指导 mRNA 合成的 DNA 链称为模板链(template strand)或称反义链(antisense strand)。贮藏在任何基因中的生物信息都必须首先被转录生成 RNA(图 3-1),才能得到表达。DNA 和 RNA 虽然很相似,只有 T 或 U 及核糖的第二位碳原子上有所不同,但它们的生物学活性却很不同。

RNA 主要以单链形式存在于生物体内，其高级结构很复杂，它们既担负着贮藏及转移遗传信息的功能，又能作为核酶直接在细胞内发挥代谢功能。

除了少数 RNA 病毒，所有 RNA 分子都来自 DNA。储存于 DNA 双链中的遗传信息通过一个被称为转录的酶促反应按照碱基互补配对的原则被转化成为单链 RNA 分子。

生物体内共有三种 RNA，即编码特定蛋白质序列的 messenger RNA (mRNA)，能特异性解读 mRNA 中的遗传信息、将其转化成相应氨基酸后加入多肽链中的 transfer RNA(tRNA)和直接参与核糖体中蛋白质合成的 ribosomal RNA(rRNA)。

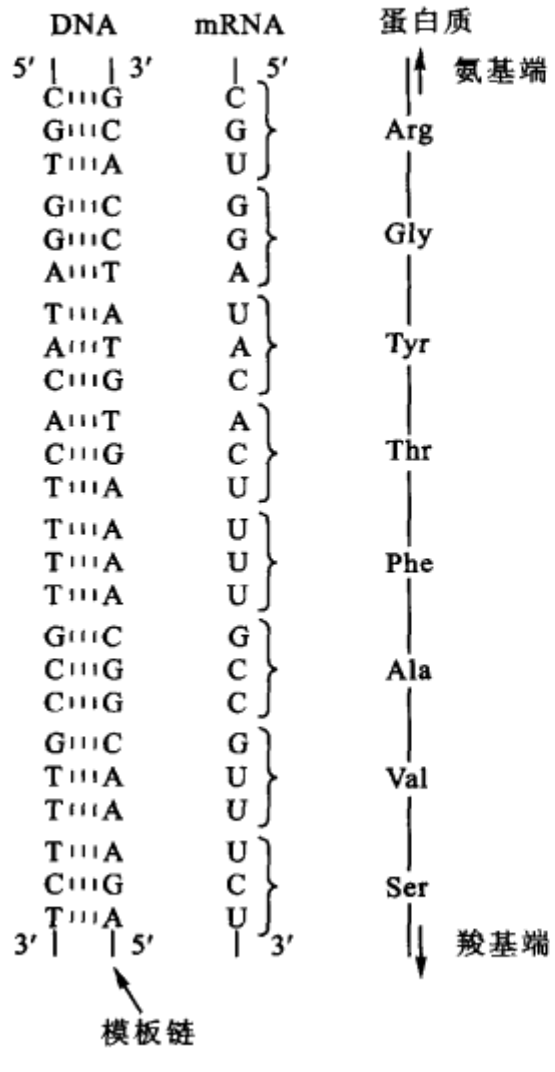



图 3-1 DNA 模板与分子及多肽链之间存在共线性关系

# 3.1 RNA 转录的基本过程

无论是原核还是真核细胞，RNA 链的合成都具有以下几个特点：RNA 是按 5′→3′方向合成的，以 DNA 双链中的反义链(模板链)为模板，在 RNA 聚合酶催化下，以 4 种三磷酸核苷(NTPs)为原料，根据碱基配对原则(A-U，T-A，G-C)，各核苷酸间通过形成磷酸二酯键相连，不需要引物的参与，合成的 RNA 带有与 DNA 编码链(有意义链)相同的序列(A-U)。转录的基本过程包括模板识别、转录起始、通过启动子及转录的延伸和终止(图 3-2)。

## 3.1.1 模板识别

模板识别(template recognition)阶段主要指 RNA 聚合酶与启动子 DNA 双链相互作用并与之相结合的过程。启动子(promoter)是基因转录起始所必需的一段 DNA 序列，是基因表达调控的上游顺式作用元件之一。

真核细胞中模板的识别与原核细胞有所不同。真核生物 RNA 聚合酶不能直接识别基因的启动子区，所以，需要一些被称为转录调控因子的辅助蛋白质按特定顺序结合于启动子上，RNA 聚合酶才能与之结合并形成复杂的转录起始前复合物(preinitiation transcription complex，PIC)，以保证有效的起

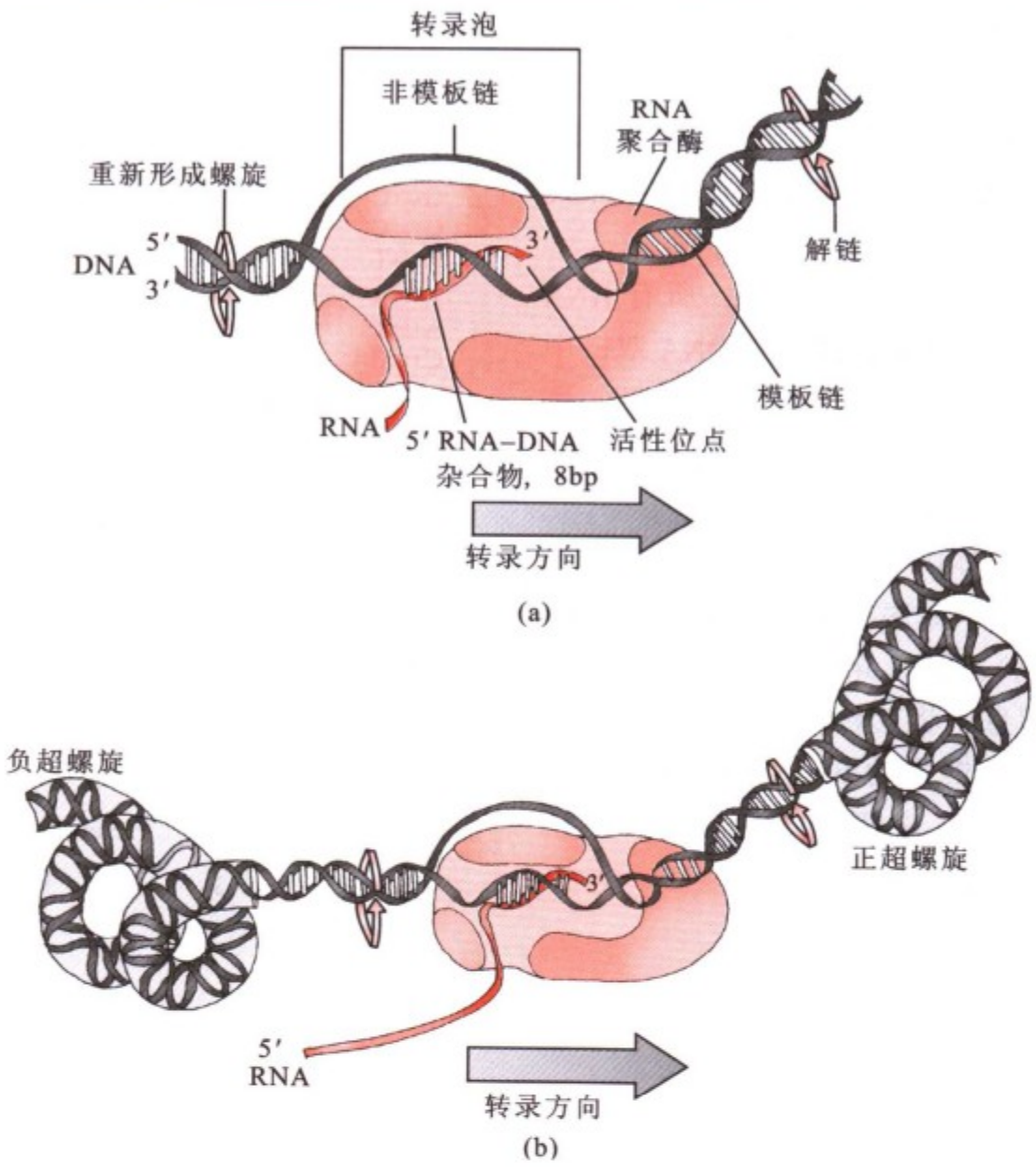



**图 3-2 大肠杆菌中依赖于 DNA 的 RNA 转录过程图示** (a) 在转录的任意时刻，转录泡中单链 DNA 的长度的 17 碱基左右，被聚合形成复合物所保护的 DNA 序列约为 35 碱基左右；(b) 由于基因转录所引起的 DNA 超螺旋结构变化。

始转录。

### 3.1.2 转录起始

转录起始(initiation)不需要引物。RNA 聚合酶结合在启动子上以后，使启动子附近的 DNA 双链解旋并解链，形成转录泡以促使底物核糖核苷酸与模板DNA 的碱基配对。转录起始就是 RNA 链上第一个核苷酸键的产生。

转录起始后直到形成 9 个核苷酸短链的过程是通过启动子阶段，此时 RNA 聚合酶一直处于启动子区，新生的 RNA 链与 DNA 模板链的结合不够牢固，很容易从 DNA 链上掉下来并导致转录重新开始。一旦 RNA 聚合酶成功地合成 9 个以上核苷酸并离开启动子区，转录就进入正常的延伸阶段。所以，通过启动子的时间代表一个启动子的强弱。一般说来，通过启动子的时间越短，该基因转录

起始的频率也越高。

### 3.1.3 转录延伸

RNA 聚合酶释放 σ 因子离开启动子后,核心酶沿模板 DNA 链移动并使新生 RNA 链不断伸长的过程就是转录的延伸(elongation)。大肠杆菌 RNA 聚合酶的活性一般为每秒 50~90 个核苷酸。随着 RNA 聚合酶的移动,DNA 双螺旋持续解开,暴露出新的单链 DNA 模板,新生 RNA 链的 3′端不断延伸,在解链区形成 RNA-DNA 杂合物。而在解链区的后面,DNA 模板链与其原先配对的非模板链重新结合成为双螺旋。

### 3.1.4 转录终止

当 RNA 链延伸到转录终止位点时,RNA 聚合酶不再形成新的磷酸二酯键,RNA-DNA 杂合物分离,转录泡瓦解,DNA 恢复成双链状态,而 RNA 聚合酶和 RNA 链都被从模板上释放出来,这就是转录的终止(termination)。

转录和翻译的速度基本相等,37℃时,转录生成 mRNA 的速度大约是每分钟 2 500 个核苷酸,即每秒钟合成 14 个密码子,而蛋白质合成的速度大约是每秒钟 15 个氨基酸。正常情况下,从一个基因开始表达到细胞中出现其 mRNA 的间隔约为 2.5 min,而再过半分钟左右就能在细胞内测到相应的蛋白质。

## 3.2 转录机器的主要成分

### 3.2.1 RNA 聚合酶

以 DNA 序列为模板的 RNA 聚合酶主要以双链 DNA 为模板(若以单链 DNA 做模板,则活性大大降低),以 4 种核苷三磷酸作为活性前体,并以 $Mg^{2+}$/$Mn^{2+}$为辅助因子,催化 RNA 链的起始、延伸和终止,它不需要任何引物,催化生成的产物是与 DNA 模板链相互补的 RNA。RNA 或 RNA-DNA 双链杂合体不能作为模板。RNA 聚合酶是转录过程中最关键的酶。每个细胞中约有 7 000 个RNA 聚合酶分子,根据细胞的生长情况,任何时候都可能有 2 000~5 000 个酶在执行转录 DNA 模板的功能。原核和真核生物的 RNA 聚合酶虽然都能催化 RNA 的合成,但在其分子组成、种类和生物化学特性上各有特色。

1. 原核生物 RNA 聚合酶

在细菌中,一种 RNA 聚合酶几乎负责所有 mRNA、rRNA 和 tRNA 的合成。大多数原核生物 RNA 聚合酶的组成是相同的,大肠杆菌 RNA 聚合酶首先由 2 个 α 亚基、一个 β 亚基、一个 β′亚基和一个 ω 亚基组成核心酶,加上一个 σ 亚基后则成为聚合酶全酶(holoenzyme),相对分子质量为 $4.65\times10^5$(图 3-3)。转录的起始过程需要全酶,由 σ 因子辨认起始点,延长过程仅需要核心酶的催化。研

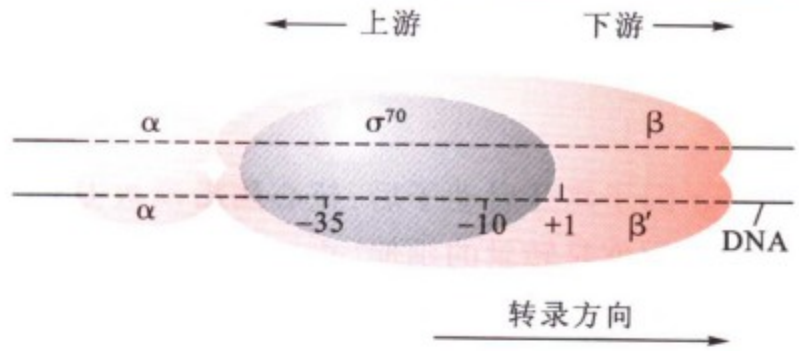



图 3-3 大肠杆菌 RNA 聚合酶的主要成分与功能分析

究发现，由 β 和 β′亚基组成了聚合酶的催化中心，它们在序列上与真核生物 RNA 聚合酶的两个大亚基有同源性。β 亚基能与模板 DNA、新生 RNA 链及核苷酸底物相结合(表 3-1)。

表 3-1 大肠杆菌 RNA 聚合酶的组成分析

| 亚基 | 基因 | 相对分子质量 | 亚基数 | 组分 | 功　能 |
|---|---|---|---|---|---|
| α | *rpoA* | $3.65\times10^4$ | 2 | 核心酶 | 核心酶组装,启动子识别。 |
| β | *rpoB* | $1.51\times10^5$ | 1 | 核心酶 | β 和 β′共同形成 RNA 合成的活性中心 |
| β′ | *rpoC* | $1.55\times10^5$ | 1 | 核心酶 | |
| | ? | $11\times10^4$ | 1 | 核心酶 | 未知 |
| σ | *rpoD* | $7.0\times10^4$ | 1 | σ 因子 | 存在多种 σ 因子,用于识别不同的启动子 |

α 亚基可能与核心酶的组装及启动子识别有关，并参与 RNA 聚合酶和部分调节因子的相互作用。实验证明,T4 噬菌体感染大肠杆菌后对 α 亚基的一个精氨酸残基进行 ADP 糖基化修饰，造成 RNA 聚合酶全酶对启动子亲和力降低。

σ 因子的作用是负责模板链的选择和转录的起始，它是酶的别构效应物，使酶专一性识别模板上的启动子。σ 因子可以极大地提高 RNA 聚合酶对启动子区 DNA 序列的亲和力,酶底结合常数提高 $10^3$ 倍,酶底复合物的半衰期可达数小时甚至数十小时。σ 因子还能使 RNA 聚合酶与模板 DNA 上非特异性位点的结合常数降低 $10^4$ 倍,使非特异性位点酶底复合物的半衰期小于 1 s。核心酶在 T7 噬菌体 DNA 上约有 1 300 个结合位点，平均结合常数为 $2\times10^{11}$。加入 σ 因子后则出现两类位点,大部分位点的结合常数在 $10^8$~$10^9$,但有 8 个位点的结合常数在 $10^{12}$~$10^{14}$,从而表明 σ 因子不仅增加聚合酶对启动子的亲和力,还降低了它对非专一位点的亲和力。

在某些细菌细胞内含有能识别不同启动子的 σ 因子,以适应不同生长发育阶段的要求,调控不同基因转录的起始。例如,枯草杆菌中就有 6 种不同相对分子质量的 σ 因子,其中 $\sigma^{55}$(右上角的数字代表相对分子质量,$\times10^3$)是枯草杆菌 RNA 聚合酶中 σ 因子的主要存在形式,出现在营养细胞中,而 $\sigma^{29}$ 则主要出现在孢子形成阶段,参与孢子形成期基因转录的调控。在大肠杆菌中,最常见的调控因子是由 *rpoD* 基因所编码的 $\sigma^{70}$,由 *rpoH* 编码的 $\sigma^{32}$ 与热休克启动子所控制

的基因转录密切相关,而由 *rpoN* 编码的 $\sigma^{54}$ 则参与细胞的氮代谢(表 3-2)。此外,由 T4 噬菌体所编码的 $\sigma^{55}$ 能与大肠杆菌 RNA 聚合酶的核心酶结合,启动 T4 晚期基因的转录。

表 3-2 大肠杆菌中的 σ 因子能识别并与启动子区的特异性序列相结合

| 因子 | 基因 | 功能 | -35 区 | 间隔(bp) | -10 区 |
|---|---|---|---|---|---|
| $\sigma^{70}$ | *rpoD* | 广泛 | TTGACA | 16 ~ 18 | TATAAT |
| $\sigma^{32}$ | *rpoH* | 热休克 | TCTCNCCCTTGAA | 13 ~ 15 | CCCCATNTA |
| $\sigma^{54}$ | *rpoN* | 氮代谢 | CTGGNA | 6 | TTGCA |

转录的起始从化学过程来看是单个核苷酸与开链启动子-酶复合物相结合构成新生 RNA 的 5′端,再以磷酸二酯键的形式与第二个核苷酸相结合,起始的终止反映在 σ 因子的释放。过去认为二核苷酸的形成就是转录起始的终止,实际上,只有当新生 RNA 链达到 6~9 个核苷酸时才能形成稳定的酶-DNA-RNA 三元复合物,才释放 σ 因子,转录进入延伸期。

与大肠杆菌相比,T3 和 T7 噬菌体的 RNA 聚合酶在结构上要简单得多,它们由一条多肽链组成,相对分子质量小于 $1\times10^5$。在 37℃下,它们的转录速度为每秒 200 个核苷酸。但是,这种简单的聚合酶只能起始存在于噬菌体中的几个启动子的转录。因此,大肠杆菌 RNA 聚合酶组成上的复杂性可能反映了它必须与大量蛋白质因子相互作用这个事实。

当 RNA 聚合酶按 5′→3′方向延伸 RNA 链时,解旋的 DNA 区域也随之移动。聚合酶可以横跨约 40 bp,而解旋的 DNA 区域大约是 17 bp。自由核苷酸能被聚合酶加到新生的 RNA 链上,并形成 DNA-RNA 杂合体。随着聚合酶在模板上的移动,靠近 3′端的 DNA 不断解旋,同时在 5′端重新形成DNA 双链,将 RNA 链挤出 DNA-RNA 杂合体。RNA 的 3′端大约有 20~30 个核苷酸与 DNA 或聚合酶相结合。

2. 真核生物 RNA 聚合酶

真核生物中共有 3 类 RNA 聚合酶,其结构比大肠杆菌 RNA 聚合酶更复杂,它们在细胞核中的位置不同,负责转录的基因不同,对 α-鹅膏蕈碱的敏感性也不同(表 3-3)。

表 3-3 真核细胞中三类 RNA 聚合酶特性比较

| 酶 | 细胞内定位 | 转录产物 | 相对活性 | 对 α-鹅膏蕈碱的敏感程度 |
|---|---|---|---|---|
| RNA 聚合酶 Ⅰ | 核仁 | rRNA | 50% ~70% | 不敏感 |
| RNA 聚合酶 Ⅱ | 核质 | hnRNA | 20% ~40% | 敏感 |
| RNA 聚合酶 Ⅲ | 核质 | tRNA | 约 10% | 存在物种特异性 |

RNA 聚合酶Ⅰ的转录产物是 45S rRNA,经剪接修饰后生成除了 5S rRNA 外的各种 rRNA。rRNA 与蛋白质组成的核糖体是蛋白质合成的场所。

RNA 聚合酶Ⅱ在核内转录生成 hnRNA，经剪接加工后生成的 mRNA 被运送到胞质中作为蛋白质合成的模板。

RNA 聚合酶Ⅲ的转录产物是 tRNA 、5S rRNA、snRNA，其中 snRNA 参与 RNA 的剪接。

α-鹅膏蕈碱是一种来自真菌毒蕈的八肽二环的剧毒物。在不同的浓度下，它对真核生物的 3 种 RNA 聚合酶有不同的作用。在动、植物及昆虫的细胞中，RNA 聚合酶Ⅱ的活性可被低浓度的 α-鹅膏蕈碱所抑制，但聚合酶Ⅰ却不受抑制。聚合酶Ⅲ对 α-鹅膏蕈碱的反应在不同的生物中有所差异，在动物细胞中高浓度的 α-鹅膏蕈可抑制转录，在昆虫中则没有抑制作用。

真核生物 RNA 聚合酶一般由 8~16 个亚基所组成(表 3-4)，相对分子质量超过 $5\times10^5$。其中，RNA 聚合酶Ⅱ的两个大亚基，RPB1 和 RPB2，与细菌核心酶的两个大亚基(β 和 β′)是同源的；其 RPB3 和 RPB11 与 α 亚基、RPB6 与 ω 亚基是同源的。虽然不同生物 3 类聚合酶的亚基种类和大小各异，它们的结构都比原核生物 RNA 聚合酶复杂，但有相似性。存在两条普遍遵循的原则：一是聚合酶中有两个相对分子质量超过 $1\times10^5$ 的大亚基；二是同种生物 3 类聚合酶有“共享”小亚基的倾向，即有几个小亚基，如 RPB6、RPC5 和 RPC9，是其中 3 类或 2 类聚合酶所共有的。

表 3-4 真核生物 RNA 聚合酶的亚基

| RNA 聚合酶Ⅰ | RNA 聚合酶Ⅱ | RNA 聚合酶Ⅲ |
|---|---|---|
| RPA 1 | RPB 1 | RPC 1 |
| RPA 2 | RPB 2 | RPC 2 |
| RPC 5 | RPB 3 | RPC 5 |
| RPC 9 | RPB 11 | RPC 9 |
| RPB 6 | RPB 6 | RPB 6 |
| 其他 9 个亚基 | 其他 7 个亚基 | 其他 11 个亚基 |

注：亚基按照相对分子质量由大到小的顺序排列。

晶体结构研究表明，无论是原核生物还是真核生物，RNA 聚合酶有着共同的结构形态，其形状大体像一只蟹爪。2006 年度诺贝尔化学奖得主罗杰·科恩伯格等通过对酵母的 RNA 聚合酶Ⅱ的晶体结构的研究全面地揭示了各亚基间的相互关系(图 3-4)。RNA 聚合酶Ⅱ分成四个主要作用模块(module)，其中包括围绕在活性中心(active center)周围的一个钳子形结构(clamp)。蟹爪的两个钳子主要是由最大的两个亚基(RPB1 和 RPB2)构成。酶的活性位点是由这两个亚基的一些区域共同组成的，位于钳子基部的一个被称为“活性中心裂隙”的区域内。只有当钳子形结构域处于开放状态时，才允许启动子 DNA 序列进入，起始基因转录。钳子形结构域延伸出来的三个平滑结构域可能在转录过程中起到促进 RNA 解旋以及帮助 DNA 重新形成螺旋的作用。

除了细胞核中的 RNA 聚合酶之外，真核生物线粒体和叶绿体中还存在着不同的 RNA 聚合酶。线粒体 RNA 聚合酶只有一条多肽链，相对分子质量小于

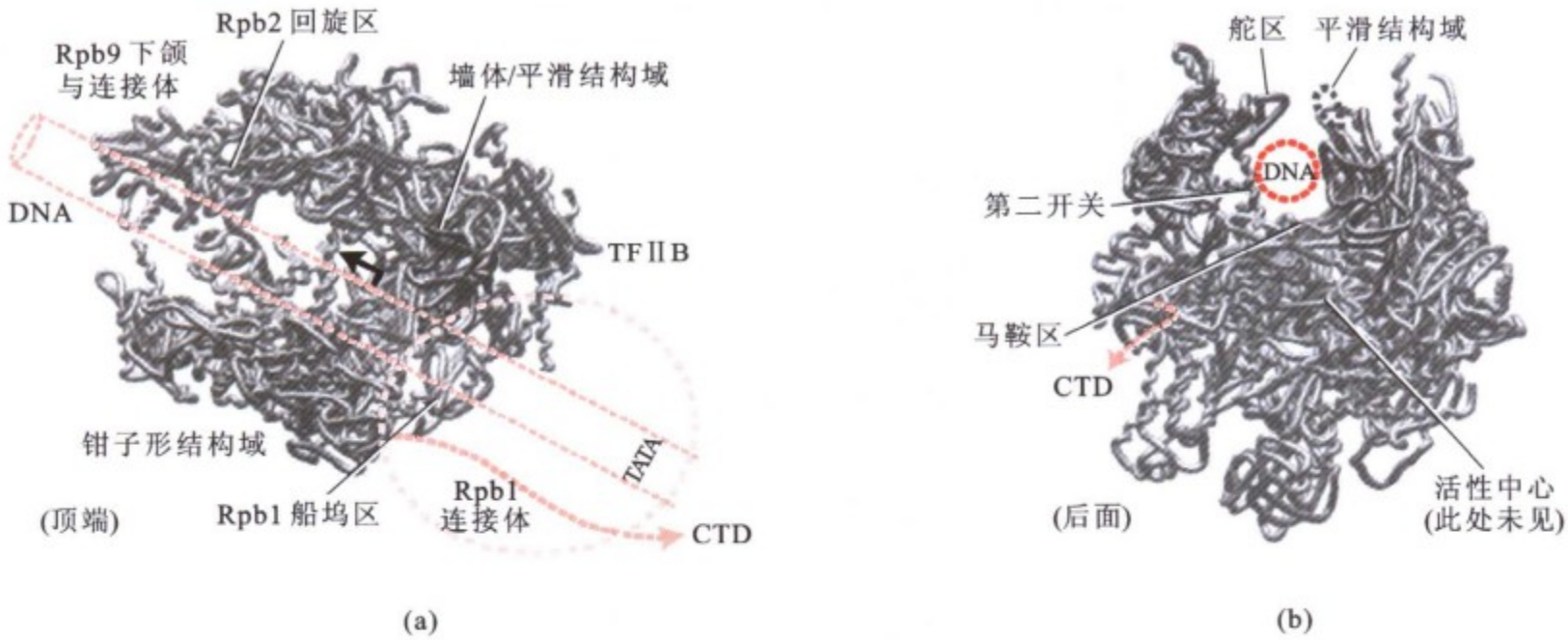



**图 3-4 模板 DNA 进入转录复合体的路径** (a)俯瞰图,双螺旋 DNA 用倾斜的圆筒表示,TF Ⅱ 转录因子的结合位点用虚圈表示;(b)后视图,DNA 从钳子形结构和墙体或平滑结构的中间穿越。

$7\times10^4$,是已知最小的 RNA 聚合酶之一,与 T7 噬菌体 RNA 聚合酶有同源性。叶绿体 RNA 聚合酶比较大,结构上与细菌中的聚合酶相似,由多个亚基组成,部分亚基由叶绿体基因组编码。线粒体和叶绿体 RNA 聚合酶活性不受 α-鹅膏蕈碱所抑制。

## 3.2.2 转录复合物

如前所述,转录可被分为 4 个阶段,即模板的识别(启动子的选择)、转录起始、RNA 链的延伸和终止。

模板的识别阶段包括 RNA 聚合酶全酶对启动子的识别,聚合酶与启动子可逆性结合形成封闭复合物(closed complex)。此时,DNA 链仍处于双链状态。伴随着 DNA 构象上的重大变化,封闭复合物转变成开放复合物 (open complex),聚合酶全酶所结合的 DNA 序列中有一小段双链被解开(图 3-5)。对于强启动子来说,从封闭复合物到开放复合物的转变是不可逆的,是快反应。开放复合物与最初的两个 NTP 相结合并在这两个核苷酸之间形成磷酸二酯键后即转变成包括 RNA 聚合酶、DNA 和新生 RNA 的三元复合物(ternary complex)。现在常用 DNA 转录循环(transcription cycle-transloction mechanism)理论(图 3-6)来解释 RNA 链的延伸。首先,三磷酸核苷酸(NTP)填补了开放的底物位点并在活性位点形成磷脂键;此时,处于 RNA 聚合酶 Ⅱ 活性位点的核酸发生“移位”,其中连接区的 α 螺旋结构从笔直变为弯折再恢复为笔直状态,为下一轮的 RNA 合成留出了空的底物位点。

除了 RNA 聚合酶之外,真核生物转录起始过程中至少还需要 7 种辅助因子参与(表 3-5),这些蛋白辅助因子统称为转录因子(transcription factor,TF)。因为不少辅助因子本身就包含多个亚基,所以转录起始复合物的相对分子质量特别大。

一般情况下,形成的三元复合物可以进入两条不同的反应途径,一是合成

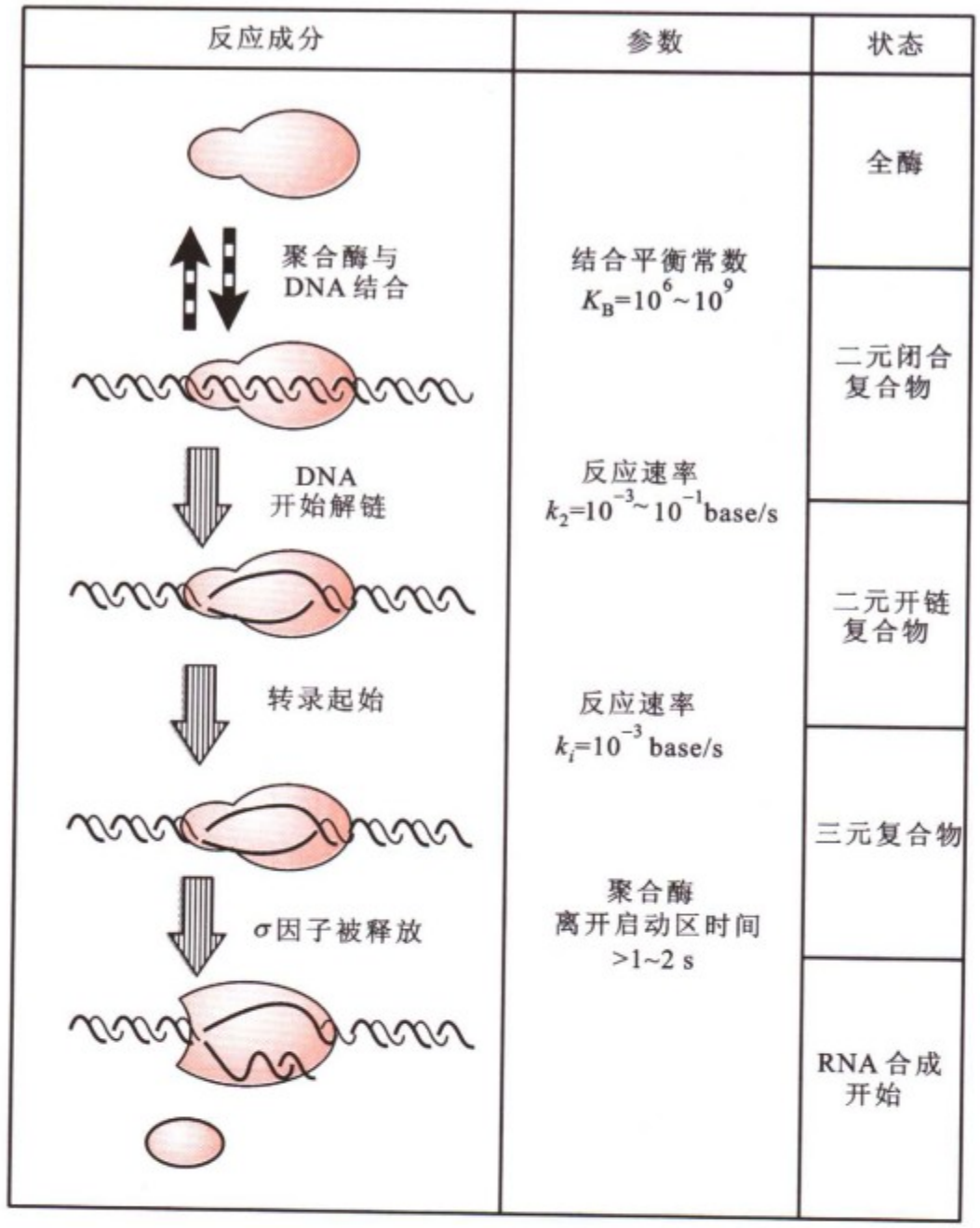



图 3-5 RNA 合成的起始

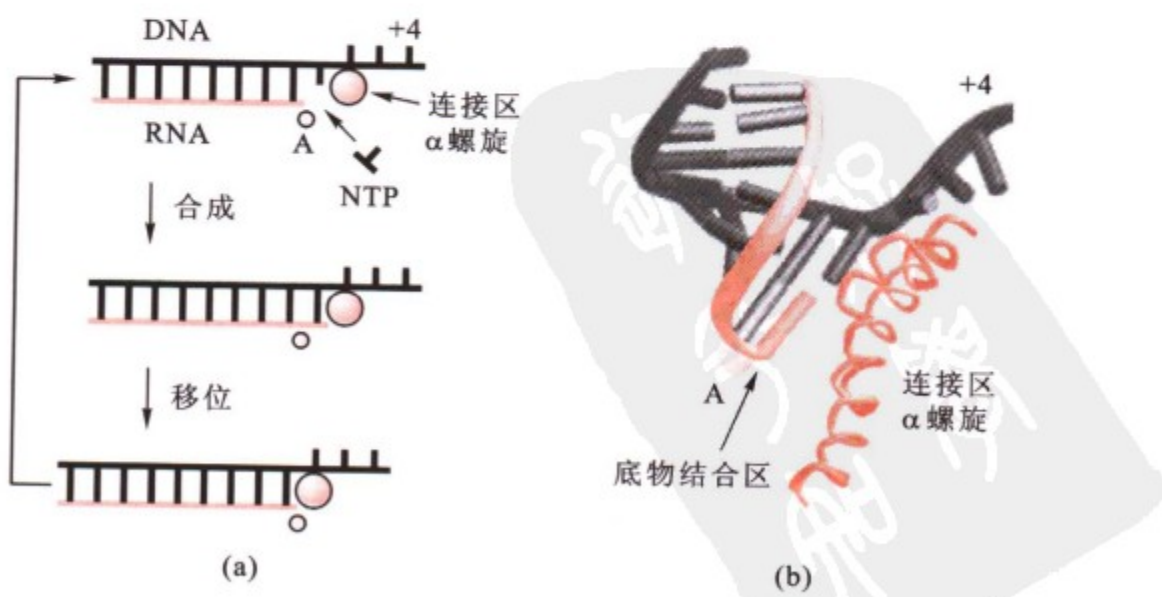



图 3-6 DNA 的转录循环假说 (a)转录循环示意图;(b)RNA 聚合酶-DNA-RNA 的转录复合物模型。

表 3-5 真核生物 RNA 聚合酶Ⅱ所形成的转录起始复合物

| 蛋白质 | 亚基数 | 亚基的相对分子质量( $\times 10^3$ ) | 功 能 |
|---|---|---|---|
| RNA 聚合酶Ⅱ | 12 | 10 ~ 220 | 催化 RNA 的生物合成 |
| TBP | 1 | 38 | 与启动子上的 TATA 区相结合 |
| TFⅡA | 3 | 12,19,35 | 使 TBP 及 TFⅡB 与启动子的结合比较稳定 |
| TFⅡB | 1 | 35 | 与 TBP 相结合,吸引 RNA 聚合酶和 TFⅡF 到启动区上 |
| TFⅡD | 12 | 15 ~ 250 | 与各种调控因子相互作用 |
| TFⅡE | 2 | 34,57 | 吸引 TFⅡH,有 ATP 酶及解链酶活性 |
| TFⅡF | 2 | 30,74 | 结合 RNA 聚合酶Ⅱ并在 TFⅡB 帮助下阻止聚合酶与非特异性 DNA 序列相结合 |
| TFⅡH | 12 | 35 ~ 89 | 在启动子区解开 DNA 双链,使 RNA 聚合酶Ⅱ磷酸化,接纳核苷酸切除修复体系 |

并释放 2~9 个核苷酸的短 RNA 转录物,即所谓的流产式起始;二是尽快释放 σ 亚基,转录起始复合物通过上游启动子区并生成由核心酶、DNA 和新生 RNA 所组成的转录延伸复合物。聚合酶全酶的作用是启动子的选择和转录的起始,而核心酶的作用是链的延伸。

转录的真实性取决于有特异的转录起始位点,转录起始后按照碱基互补原则准确地转录模板 DNA 序列及具有特异的终止部位。RNA 的合成是在模板 DNA 的启动子位点上起始的,而这个任务是靠 σ 因子来完成的。RNA 聚合酶的核心酶虽可合成 RNA,但不能找到模板 DNA 上的起始位点,所以核心酶的产物是不均一的,而且 DNA 的两条链都可作为核心酶的模板。只有带 σ 因子的全酶才能专一地与 DNA 上的启动子结合,选择其中一条链作为模板,合成 RNA 链。σ 因子的作用在于帮助转录起始,一旦转录开始,它就脱离了起始复合物,而由核心酶负责 RNA 链的延伸。

转录延伸复合物是转录循环中一个十分重要的环节。与转录起始复合物相比,延伸复合物极为稳定,可以长时间地与 DNA 模板相结合而不解离。只有在它遇到转录终止信号时,RNA 聚合酶才停止加入新的核苷酸,RNA-DNA 复合物解离,释放转录产物并导致聚合酶本身从模板 DNA 上解离下来。

## 3.3 启动子与转录起始

大肠杆菌 RNA 聚合酶与启动子的相互作用主要包括启动子区的识别、酶与启动子的结合及 σ 因子的结合与解离。

### 3.3.1 启动子区的基本结构

启动子是一段位于结构基因 5′端上游区的 DNA 序列，能活化 RNA 聚合

酶,使之与模板 DNA 准确地结合并具有转录起始的特异性。因为基因的特异性转录取决于酶与启动子能否有效地形成二元复合物，所以,RNA 聚合酶如何有效地找到启动子并与之相结合是转录起始过程中首先要解决的问题。有实验表明，对许多启动子来说,RNA 聚合酶与之相结合的速率至少比布朗运动中的随机碰撞高 100 倍。

现在我们已经知道,转录的起始是基因表达的关键阶段,而这一阶段的重要问题是 RNA 聚合酶与启动子的相互作用。启动子的结构影响了它与 RNA 聚合酶的亲和力,从而影响了基因表达的水平。

转录单元(transcription unit)是一段从启动子开始至终止子(terminator)结束的 DNA 序列,RNA 聚合酶从转录起始位点开始沿着模板前进,直到终止子为止,转录出一条 RNA 链(图 3-7)。在细菌中,一个转录单元可以是一个基因,也可以是几个基因。转录起始位点是指与新生 RNA 链第一个核苷酸相对应 DNA 链上的碱基,研究证实通常为一个嘌呤。常把起点前面,即 5′端的序列称为上游序列(upstream),而把其后面即 3′端的序列称为下游序列(downstream)。在描述碱基的位置时,一般用数字表示,起点为+1,下游方向依次为+2,+3,…,上游方向依次为-1,-2,-3……序列的书写方向通常是固定的,使转录从左(上游)向右(下游)进行,mRNA 同样按照 5′→3′方向书写。

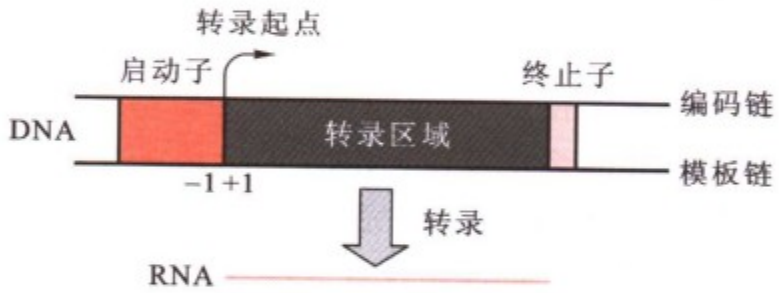



图 3-7 RNA 的转录单元示意图

启动子区是 RNA 聚合酶的结合区,其结构直接关系到转录的效率。那么,启动子区有什么结构特点呢?

Pribnow 设计了一个实验，他把 RNA 聚合酶全酶与模板 DNA 结合后，用 DNase Ⅰ水解 DNA,然后用酚抽提,沉淀纯化 DNA 后得到一个被 RNA 聚合酶保护的 DNA 片段（图 3-8),约有 41~44 个核苷酸对。他先后分离了 fd 噬菌体、T7 噬菌体的 $A_2$ 及 $A_3$ 启动子、λ 噬菌体的 $P_R$ 启动子及大肠杆菌乳糖操纵子的 UV5 启动子等 5 段被酶保护的区域,并进行了序列分析。以后又有人做了 50 多个启动子的序列分析后发现，在被保护区内有一个由 5 个核苷酸组成的共同序列,是 RNA 聚合酶的紧密结合点,现在称为 Pribnow 区(Pribnow box),这个区的中央大

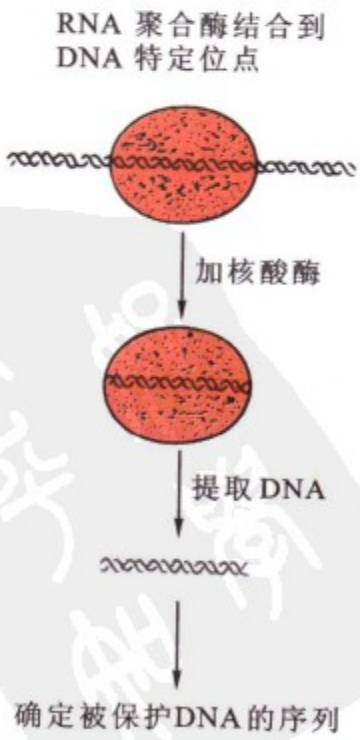



图 3-8 转录启动子区 DNA 序列的分离纯化过程图示

约位于起点上游 10 bp 处,所以又称为-10 区。

提纯被保护的片段后却发现,RNA 聚合酶并不能重新结合或并不能选择正确的起始点，表明在保护区外可能还存在与 RNA 聚合酶对启动子的识别有关的序列。果然,科学家不久就从噬菌体的左、右启动子 $P_L$ 及 $P_R$ 和 SV40 启动子的-35 bp 附近找到了另一段共同序列:TTGACA。经过数年的努力，分析了 46 个大肠杆菌启动子的序列以后确证绝大部分启动子都存在这两段共同序列,即位于-10 bp 处的 TATA 区和-35 bp 处的 TTGACA 区（图 3-9)。现已查明,-10 位的 TATA 区和-35 位的 TTGACA 区是 RNA 聚合酶与启动子的结合位点,能与 σ 因子相互识别而具有很高的亲和力。

-35 区　　-10 区

$\cdots T_{85}T_{83}G_{81}A_{61}C_{69}A_{52}\cdots T_{89}A_{89}T_{50}A_{65}A_{100}\cdots$

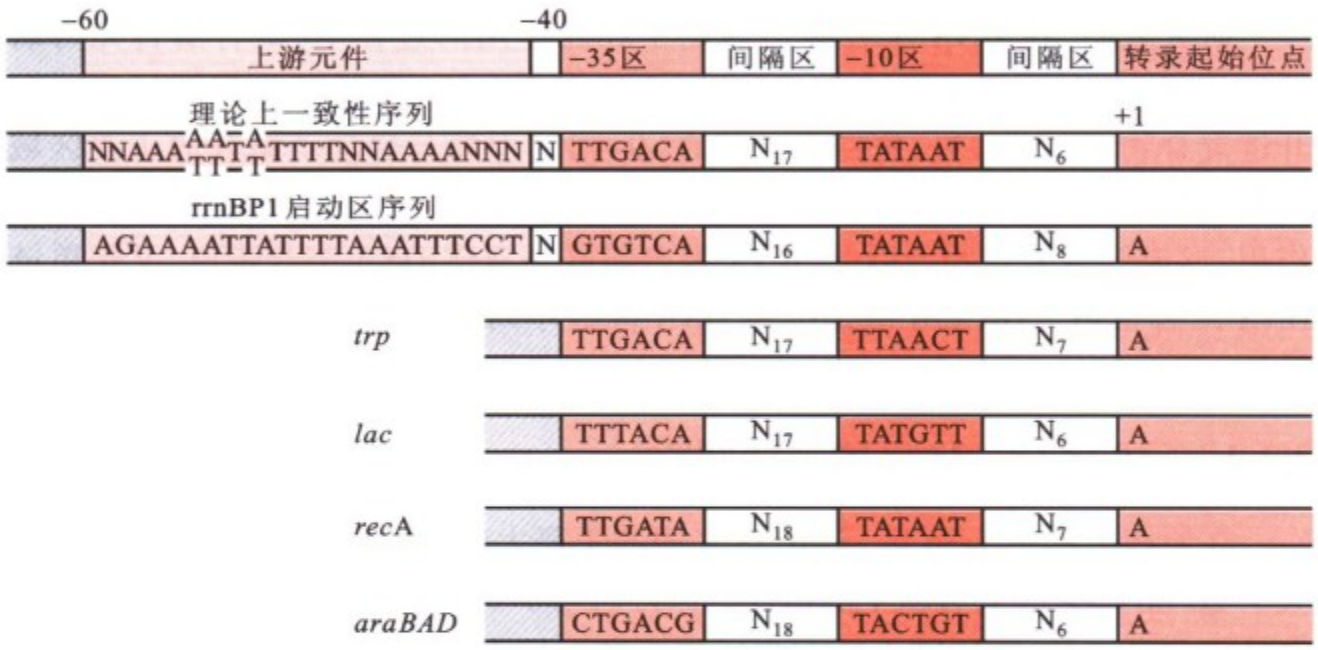



图 3-9　大肠杆菌 RNA 聚合酶全酶所识别的启动子区

在真核生物基因中,Hogness 等先在珠蛋白基因中发现了类似 Pribnow 区的 Hogness 区(Hogness box),这是位于转录起始点上游-30~-25 bp 处的共同序列 TATAAA,也称为 TATA 区(图 3-10)。另外,在起始位点上游-78~-70 bp 处还有另一段共同序列 CCAAT，这是与原核生物中-35 bp 区相对应的序列，称为 CAAT 区(CAAT box)。

图 3-10　真核生物 RNA 聚合酶Ⅱ所识别的启动子区

## 3.3.2　启动子区的识别

一般认为,RNA 聚合酶并不直接识别碱基对本身，而是通过氢键互补的方

式加以识别。在启动子区 DNA 双螺旋结构中,腺嘌呤分子上的 $N^6$、鸟嘌呤分子上的 $N^2$、胞嘧啶分子上的 $N^4$ 都是氢键供体,而腺嘌呤分子上的 $N^7$、$N^3$,胸腺嘧啶分子上的 $O^4$、$O^2$, 鸟嘌呤分子上的 $N^7$、$O^6$、$N^3$ 和胞嘧啶分子上的 $O^2$ 都是氢键受体。由于它们分别处于 DNA 双螺旋的大沟或小沟内,因此都具有特定的方位,而酶分子中也有处于特定空间构象的氢键受体与供体,当它们与启动子中对应的分子在一定距离内互补时,就形成氢键,相互结合。这种氢键互补学说较为圆满地解释了启动子功能既受 DNA 序列影响,又受其构象影响这一事实。当保守区某些碱基的取代不影响 DNA 上氢键的方位和特性时, 启动子原有的功能保持不变;当局部 DNA 构象或电荷密度改变影响了这些基团的相对方位时,启动子的功能就会受到影响。

### 3.3.3 RNA 聚合酶与启动子区的结合

研究表明,在 RNA 聚合酶与启动子相互作用的过程中,聚合酶首先与启动子区闭合双链 DNA 相结合,形成二元闭合复合物,然后经过解链得到二元开链复合物(图 3-11)。解链区一般在-9~+13,而酶与启动子结合的主要区域在其上游。一旦开链区解链,酶分子能以正确的取向与解链后的有关单链相互作用,形成开链复合物。因此,RNA 聚合酶既是双链 DNA 结合蛋白,又是单链 DNA 结合蛋白。DNA 开链是按照 DNA 模板序列正确引入核苷酸底物的必要条件。

### 3.3.4 -10 区与 -35 区的最佳间距

在原核生物中,-35 区与-10 区之间的距离大约是 16~19 bp,小于 15 bp 或大于 20 bp 都会降低启动子的活性。保持启动子这两段序列以及它们之间的距离是十分重要的,否则就会改变它所控制基因的表达水平。这可以被解释为一旦-35 区相对于-10 区旋转所产生的超螺旋发生改变(增减 1 bp 就会使两者之间的夹角发生 36°的变化),若要使酶与 DNA 在这个区域内保持正确的取向,就必须使二者之一发生扭曲,需要增加结合自由能。

在细菌中常见两种启动子突变,一种叫下降突变(down mutation),如果把 Pribnow 区从 TATAAT 变成 AATAAT 就会大大降低其结构基因的转录水平;另一类突变叫上升突变(up mutation),即增加 Pribnow 区共同序列的同一性。例如,在乳糖操纵子的启动子中,将其 Pribnow 区从 TATGTT 变成 TATATT,就会提高启动子的效率,提高乳糖操纵子基因的转录水平。

### 3.3.5 增强子及其功能

除了启动子以外,近年来发现还有另一序列与转录的起始有关。在 SV40 的转录单元上发现它的转录起始位点上游约 200 bp 处有两段 72 bp 长的重复序列,它们不是启动子的一部分,但能增强或促进转录的起始,除去这两段序列会大大降低这些基因的转录水平, 若保留其中一段或将之取出插至 DNA 分子的任何部位,就能保持基因的正常转录。因此,称这种能强化转录起始的序列为增

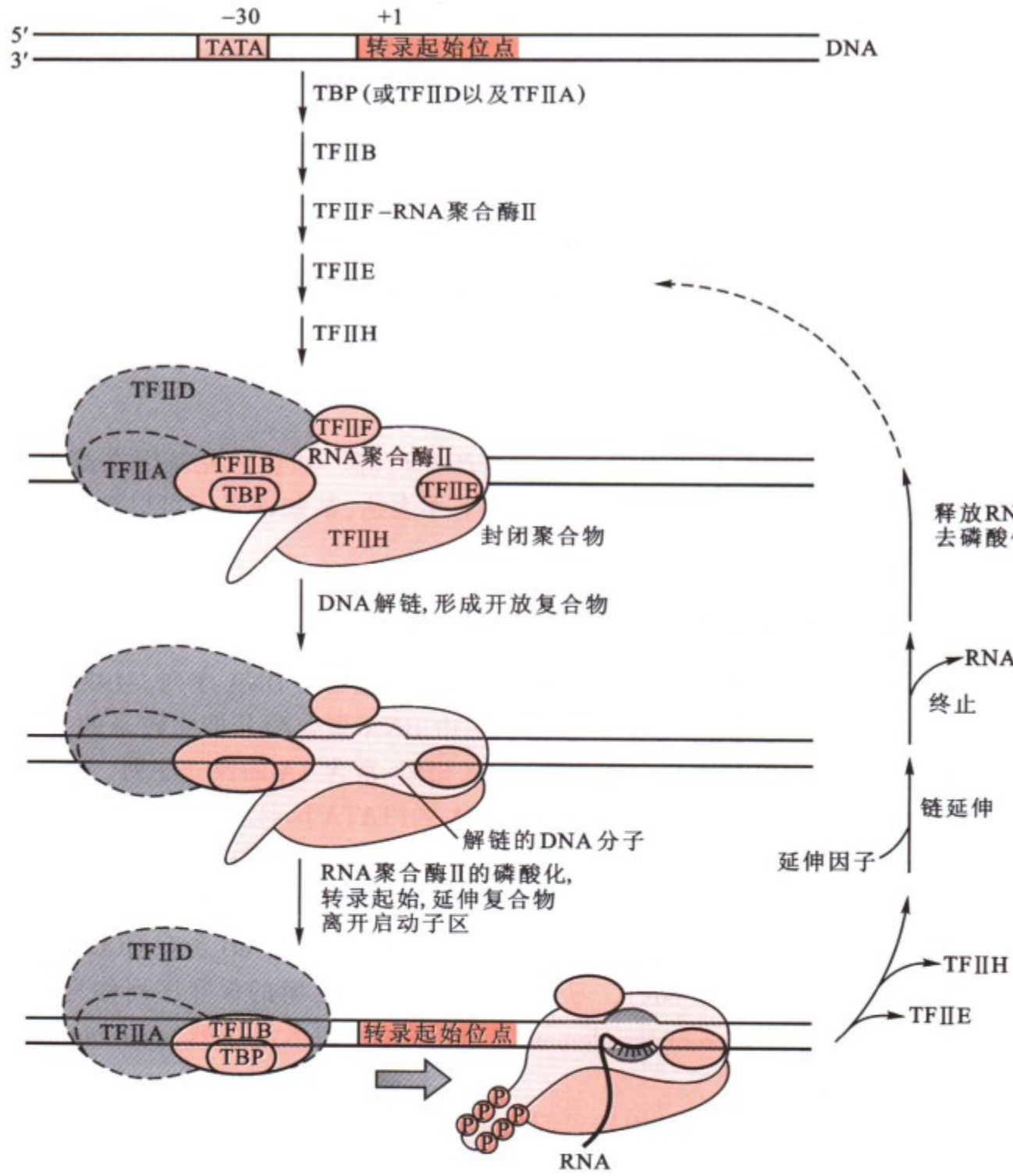



图 3-11 由 RNA 聚合酶Ⅱ指导的基因转录过程图示

强子或强化子(enhancer)。除 SV40 外,还在反转录病毒基因、免疫球蛋白基因、胰岛素基因、胰糜蛋白酶基因等许多基因的启动区中陆续发现了增强子的存在。

有人曾把β-珠蛋白基因置于带有上述 72 bp 序列的 DNA 分子上,发现它在体内的转录水平提高了 200 倍。而且,无论把这段 72 bp 序列放在转录起点上游 1 400 bp 处还是下游 3 300 bp 处,它都有转录增强作用。增强子很可能通过影响染色质 DNA-蛋白质结构或改变超螺旋的密度而改变模板的整体结构,从而使得 RNA 聚合酶更容易与模板 DNA 相结合,起始基因转录。

增强子具有下列特点:

① 远距离效应 一般位于上游-200 bp 处,但可增强远处启动子的转录,

即使相距>10 kb 也能发挥作用。

② 无方向性　无论位于靶基因的上游、下游或内部都可发挥增强转录的作用。

③ 顺式调节　只调节位于同一染色体上的靶基因，而对其他染色体上的基因没有作用。

④ 无物种和基因的特异性　可以连接到异源基因上发挥作用。

⑤ 具有组织特异性　SV40 的增强子在 3T3 细胞中比多瘤病毒的增强子要弱,但在 Hela 细胞中 SV40 的增强子比多瘤病毒的要强 5 倍。增强子的效应需特定的蛋白质因子参与。

⑥ 有相位性　其作用和 DNA 的构象有关。

⑦ 有的增强子可以对外部信号产生反应　如热休克基因在高温下才表达,金属硫蛋白基因在镉和锌的存在下才表达,某些增强子可以被固醇类激素所激活。

### 3.3.6　真核生物启动子对转录的影响

所谓启动子是指确保转录精确而有效地起始的 DNA 序列。1979 年美国科学家 Goldberg 首先注意到真核生物中由 RNA 聚合酶Ⅱ催化转录的 DNA 序列 5′上游区有一段与原核生物 Pribnow 区相似、富含 TA 的保守序列。由于该序列前 4 个碱基为 TATA,所以又称为 TATA 区(TATA box)。此后 10 多年间,科学家通过对许多基因启动子区的分析,发现绝大多数功能蛋白基因的启动子都具有共同的结构模式。简单地说,真核基因的启动子在-35~-25 区含有 TATA 序列,在-80~-70 区含有 CCAAT 序列(CAAT box),在-110~-80 含有 GCCACACCC 或 GGGCGGG 序列(GC box)。习惯上,将 TATA 区上游的保守序列称为上游启动子元件(upstream promoter element,UPE)或称上游激活序列(upstream activating sequence,UAS)。

比较原核和真核基因转录起始位点上游区的结构,发现两者之间存在很大的差别 (图 3-12)。原核基因启动区范围较小，一般情况下,TATAAT(Pribnow 区)的中心位于-10~-7,上游-70~-30 区为正调控因子结合序列,-20~+1 区为负调控因子结合序列。真核基因的调控区较大,TATAA/TA 区位于-30~-20,而-110~-40 区为上游激活区。除 Pribnow 区之外，原核基因启动子上游只有 TTGACA 区(-40~-30)作为 RNA 聚合酶的主要结合位点,参与转录调控;而真核基因除了含有可与之相对应的 CAAT 区之外，大多数基因还拥有 GC 区和增强子区。

TATA 区和上游启动子元件(CAAT 区或 GC 区)的作用有所不同。前者的主要作用是使转录精确地起始,如果除去 TATA 区或进行碱基突变,转录产物下降的相对值不如 CAAT 区或 GC 区突变后明显,但发现所获得的 RNA 产物起始点不固定。研究 SV40 晚期基因启动子发现,上游激活区的存在与否,对该启动子的生物活性有着根本性的影响。若将该基因 5′上游-47~-21 核苷酸序列切除,基因完全不表达(图 3-13)。

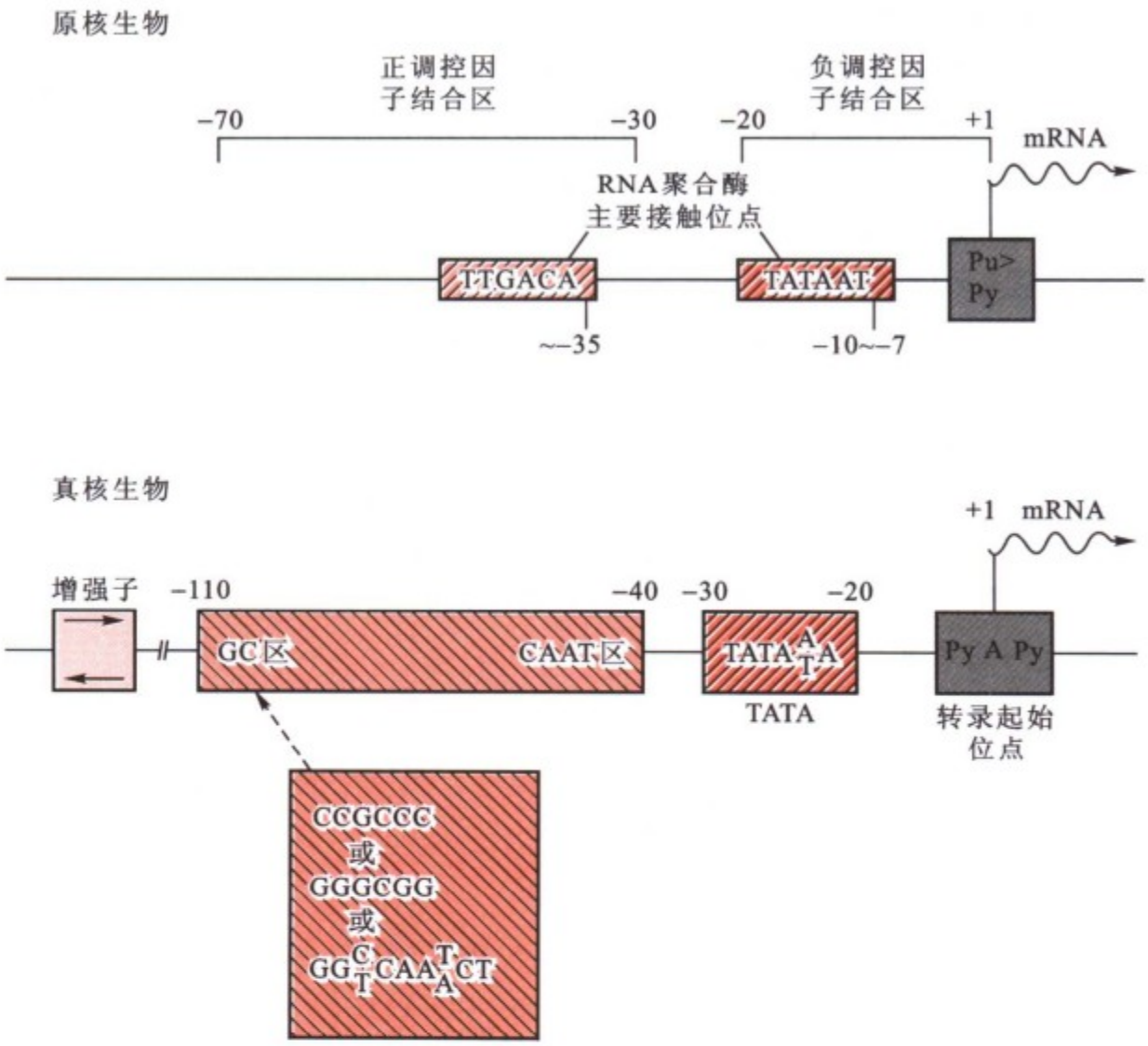



图 3-12 原核和真核生物基因转录起始位点上游区的结构比较

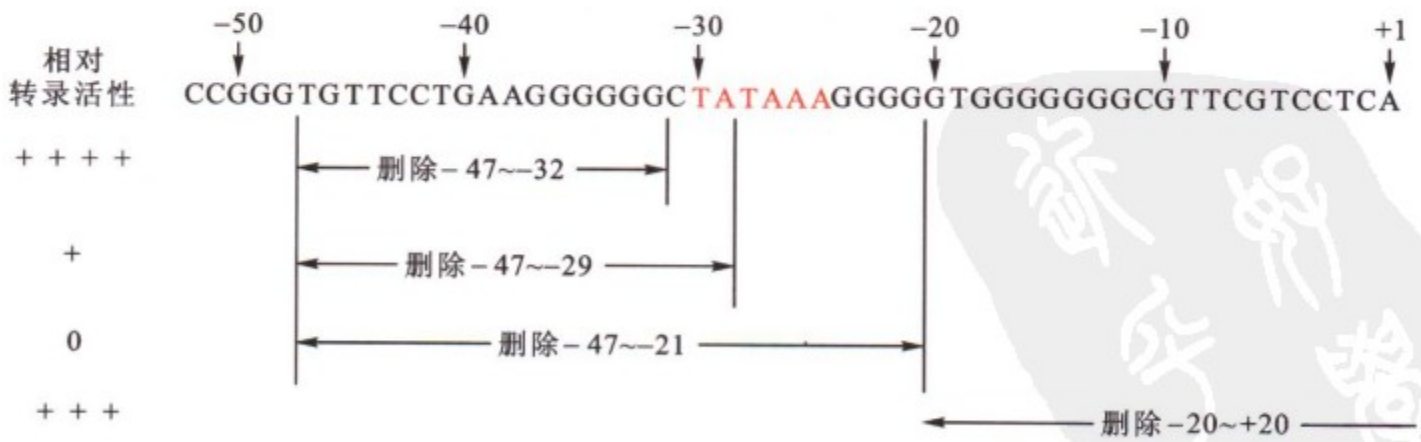



图 3-13 SV40 基因启动子上 TATAAA 及邻近区域对基因转录活性的影响

CAAT 区和 GC 区主要控制转录起始频率,基本不参与起始位点的确定。CAAT 区对转录起始频率的影响最大,该区任意碱基的改变都将极大地影响靶基因的转录强度,而启动区其他序列中一两个碱基的置换则没有太大的影响。此外,在 TATA 区和相邻的 UPE 区之间插入核苷酸也会使转录减弱,在人类β-血红蛋白基因启动区的两个 UPE 序列之间插入 15 个核苷酸,该启动子完全失去功能。CAAT 和 GC 区相对于 TATA 区的取向(5′→3′或 3′→5′),对转录强度的影响不大。

尽管这 3 种保守序列都有着重要功能,但并不是每个基因的启动子区都包含这 3 种序列。有些基因,如 SV40 的早期基因,缺少 TATA 和 CAAT 区,只有 6 个串联在上游-110~-40 位点的 GC 区;有些基因,如组蛋白 $H_2B$,不含 GC 区,但有两个 CAAT 区,一个 TATA 区。研究发现,真核细胞中存在着大量特异性或组成型表达的、能够与不同基因启动子区 UPE 相结合的转录调控因子(表 3-6)。基因转录实际上是 RNA 聚合酶、转录调控因子和启动子区各种调控元件相互作用的结果。

表 3-6 真核细胞中的部分转录调控因子分析

| 转录因子 | DNA 结合区的特征 | DNA 结合区序列 | 在生物体内的分布 |
|---|---|---|---|
| **组成型转录因子** | | | |
| CTF/NF1 | CAAT 区 | 5′- GCCAATCT - 3′ | 广泛 |
| CP | CAAT 区 | 5′- GCCAATCT - 3′ | 广泛 |
| C/EBP | CAAT 区 | 5′- GCCAATCT - 3′ | 广泛 |
| Sp1 | GC 区 | 5′- GGGCGG - 3′ | 广泛 |
| Oct - 1 | 八碱基区 | 5′- ATGCAAAT - 3′ | 广泛 |
| Oct - 2 | 八碱基区 | 5′- ATGCAAAT - 3′ | B 淋巴细胞 |
| **应答蛋白** | | | |
| 热激因子 | HSE | 5′- CNNGAANNTCCNNG - 3′ | 受热激以后 |
| **血清反应因子** | SRE | 5′- CCATATTAGG - 3′ | 受血清生长因子刺激 |
| **细胞特异性因子** | | | |
| GATA - 1 | — | 5′- GATA - 3′ | 成红细胞 |
| Pit - 1 | — | 5′- ATATTCAT - 3′ | 腺体细胞 |
| MyoD1 | E 区 | 5′- CANNG - 3′ | 成肌细胞 |
| NF - κB | κB 位点 | 5′- GGGACTTTCC - 3′ | 淋巴细胞 |

### 3.3.7 转录的抑制

RNA 转录的抑制剂根据其作用的性质主要可以分为两大类:第一类是 DNA 模板功能抑制剂,通过与 DNA 结合而改变模板的功能;第二类是 RNA 聚合酶的抑制物,它们与 RNA 聚合酶结合而抑制其活力。表 3-7 列举了一些典型

表 3-7 常见的转录抑制剂

| 抑制剂 | 靶 酶 | 抑制作用 |
|---|---|---|
| 利福霉素 | 细菌全酶 | 和 β 亚基结合,抑制起始 |
| 利迪链霉素 | 细菌核心酶 | 和 β 亚基结合,抑制起始 |
| 放射线素 D | 真核 RNA 聚合酶 I | 与 DNA 结合,阻止延伸 |
| α-鹅膏蕈碱 | 真核 RNA 聚合酶 II | 与 RNA 聚合酶 II 结合 |

的 RNA 转录的抑制剂。放线菌素 D(actinomycin D)是 DNA 模板功能的抑制物,可与 DNA 形成非共价复合物,抑制其作为模板的功能。1 mmol/L 的放线菌素 D 就可有效地抑制转录过程,它具有抗菌和抗癌作用。

利福霉素(rifamycin)、利迪链霉素(streptolydigin)和 α-鹅膏蕈碱都是 RNA 聚合酶的抑制物。其中,利福霉素能强烈地抑制革兰阳性菌和结合杆菌,具有广谱抗菌作用,能特异性地抑制细菌 RNA 聚合酶的活性。利迪链霉素与细菌 RNA 聚合酶的 β 亚基结合,抑制转录的起始。α-鹅膏蕈碱主要抑制真核生物的 RNA 聚合酶。

除了上述抑制剂外,还有一些嘌呤和嘧啶类似物,如 5-氟尿嘧啶、6-巯基嘌呤、硫鸟嘌呤、8-氮鸟嘌呤以及 6-氮尿嘧啶等,可以作为核苷酸代谢拮抗物来抑制核酸前体的合成。

## 3.4 原核与真核生物 mRNA 的特征比较

信使核糖核酸通常用 mRNA 表示,是英文 messenger RNA 的简称,在大肠杆菌细胞内占总 RNA 的 2%左右 (tRNA 占 16%, 而 rRNA 则占 80%以上)。尽管科学家很早就怀疑生物细胞内存在能将遗传信息从 DNA 上转移到蛋白质分子上的信使(或称模板),但由于 mRNA 在细菌细胞内的半衰期很短,科学家直到 20 世纪 70 年代初才首次将这一重要物质从细胞中分离出来。现已查明,许多基因的 mRNA 在体内还是相当稳定的,半衰期从几个小时到几天不等。目前,人们已能从几乎所有生物体内分离纯化编码任何蛋白质的 mRNA。

虽然 mRNA 在所有细胞内执行着相同的功能,即通过密码三联子翻译生成蛋白质,其生物合成的具体过程和成熟 mRNA 的结构在原核和真核细胞内是不同的。

真核细胞 mRNA 的最大特点在于它往往以一个较大相对分子质量的前体 RNA 形式出现在核内,需要经过转录后加工。只有成熟的、相对分子质量明显变小并经化学修饰的 mRNA 才能进入细胞质,参与蛋白质的合成。所以,真核细胞 mRNA 的合成和功能表达发生在不同的空间和时间范畴内。

原核生物中,mRNA 的转录和翻译不仅发生在同一个细胞空间里, 而且这两个过程几乎是同步进行的, 蛋白质合成往往在 mRNA 刚开始转录时就

被引发了。一个原核细胞的 mRNA(包括病毒)有时可以编码几个多肽,而一个真核细胞的 mRNA 最多只能编码一个多肽。原核生物常以 AUG (有时 GUG,甚至 UUG)作为起始密码子,而真核生物几乎永远以 AUG 作为起始密码子。

### 3.4.1 原核生物 mRNA 的特征

1. 原核生物 mRNA 的半衰期短

细菌基因的转录与翻译是紧密相连的,基因转录一旦开始,核糖体就结合到新生 mRNA 链的 5′端,启动蛋白质合成,而此时该 mRNA 的 3′端还远远没有转录完全。在电子显微镜下,我们可以看到一连串核糖体紧紧跟在 RNA 聚合酶后面。

绝大多数细菌 mRNA 的半衰期短得令人吃惊,mRNA 的降解紧随着蛋白质翻译过程发生。现在一般认为,转录开始 1 min 后,降解就开始了,这就是说,当一个 mRNA 的 5′端开始降解时,其 3′端部分可能仍在合成或被翻译。mRNA 降解的速度大概只有转录或翻译速度的一半。科学家发现每过大约 2 min,体系中出现新生蛋白质的速度就下降 50%。

因为基因转录和多肽链延伸的速度基本相同, 所以细胞内某一基因产物(蛋白质)的多少决定于转录和翻译的起始效率,以大肠杆菌色氨酸合成酶基因为例,平均每分钟约有 15 次转录起始,这些 mRNA 链从生成到降解平均被翻译 10 次,所以,稳定状态下细胞中每分钟生成 150 个多肽。在讨论 mRNA 半衰期的时候我们不应忘记, 所有这一切都只是统计学上的平均值。事实上, 新生 mRNA 与成熟 mRNA 受核酸酶“攻击”的概率是相等的。所以,有些 mRNA 链可能被翻译了许多次, 而同一群体中的另外一些 mRNA 分子可能根本没有被翻译。虽然 mRNA 的寿命具有随机性,但某一 mRNA 的翻译产量却是大致稳定的。

2. 许多原核生物 mRNA 可能以多顺反子的形式存在

细菌 mRNA 可以同时编码不同的蛋白质, 我们把只编码一个蛋白质的 mRNA 称为单顺反子 mRNA (monocistronic mRNA), 把编码多个蛋白质的 mRNA 称为多顺反子 mRNA(polycistronic mRNA)。多顺反子 mRNA 是一组相邻或相互重叠基因的转录产物,这样的一组基因可被称为一个操纵子(operon),是生物体内的重要遗传单位,如大肠杆菌乳糖操纵子转录成编码 3 条多肽的多顺反子 mRNA,经过翻译生成 β-半乳糖苷酶、β-半乳糖苷透过酶及 β-半乳糖苷乙酰基转移酶。

几乎所有 mRNA 都可以被分成 3 部分:编码区和位于 AUG 之前的 5′端上游非编码区以及位于终止密码子之后不翻译的 3′端下游非编码区。编码区从起始密码子 AUG 开始经一连串编码氨基酸的密码子直至终止密码子。对于第一个顺反子来说,一旦 mRNA 的 5′端被合成,翻译起始位点即可与核糖体相结合, 而后面几个顺反子翻译的起始就会受其上游顺反子结构的调控。

一种情况是，第一个蛋白质合成终止以后，核糖体分解成大、小亚基，脱离 mRNA 模板，第二个蛋白的翻译必须等到新的小亚基和大亚基与该蛋白起始密码子相结合后才可能开始(图 3-14a)。当然，前一个多肽翻译完成以后，核糖体大、小亚基分离，小亚基也可能不离开 mRNA 模板，而是迅速与游离的大亚基结合，启动第二个多肽的合成(图 3-14b)。在噬菌体 RNA 中，一个顺反子的翻译有时完全取决于它前面顺反子的翻译，因为噬菌体 RNA 往往形成复杂的二级结构，只有第一个翻译起始位点是暴露的，在这个顺反子翻译产生多肽的过程中，核糖体的运动破坏了后续顺反子的二级结构，使起始位点较容易与核糖体相结合形成起复合物(图 3-15)。

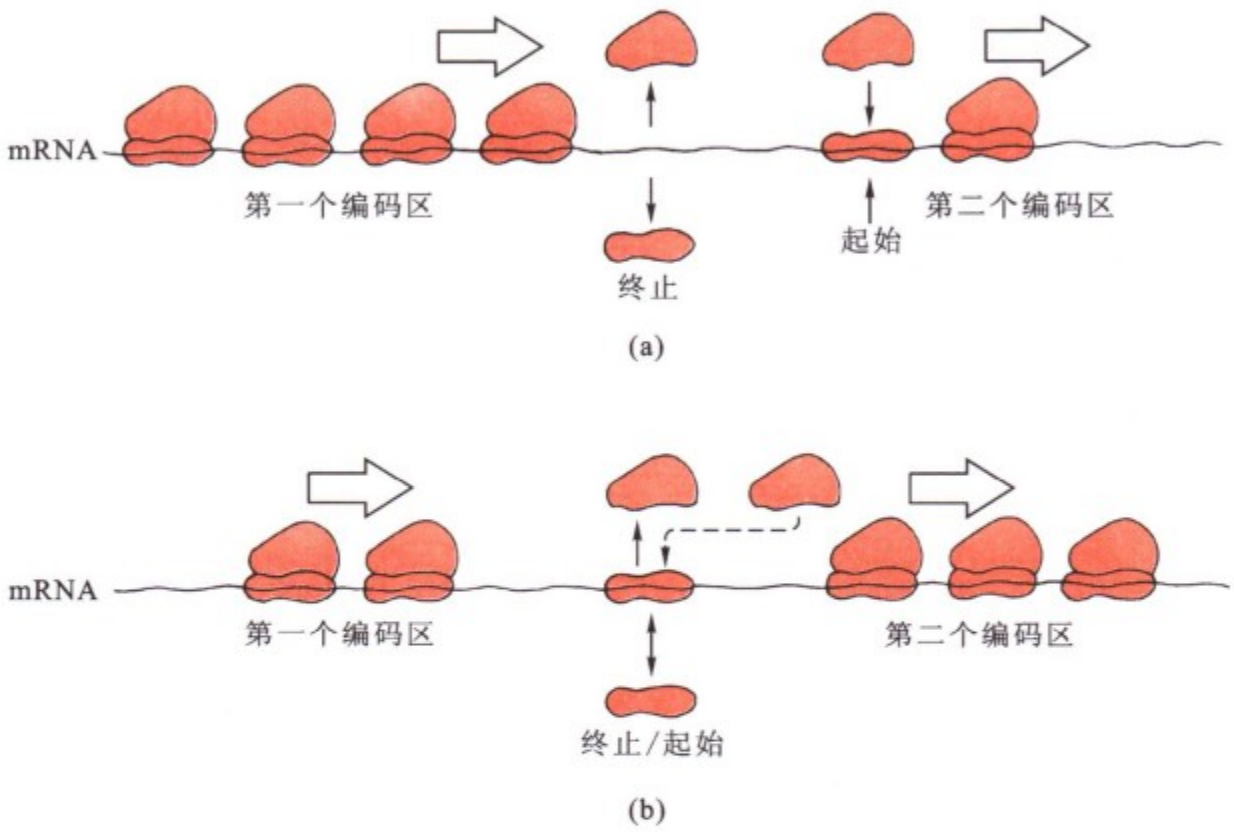
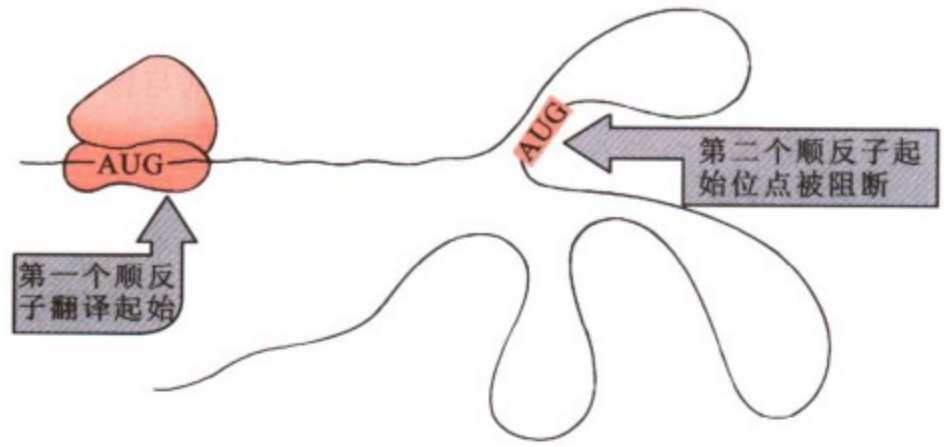



**图 3-14 原核生物多顺反子基因中后续编码区翻译起始受顺反子之间距离的影响**

(a) 当两个顺反子之间距离较远时，前一顺反子翻译的终止与后一顺反子翻译的起始是相互独立的；(b) 当两个顺反子之间距离较近时，前一顺反子翻译的终止与后一顺反子翻译的起始相衔接，30S 小亚基始终与 mRNA 结合，只有 50S 大亚基可能与 mRNA 分离。

3. 原核生物 mRNA 的 5′端无帽子结构，3′端没有或只有较短的 poly A 结构

原核生物起始密码子 AUG 上游 7~12 个核苷酸处有一被称为 SD 序列(Shine-Dalgarno sequence)的保守区，因为该序列与 16S rRNA 3′端反向互补，所以被认为在核糖体-mRNA 的结合过程中起作用。如果把 16S rRNA 3′端的 6 个保守核苷酸反过来写，则为 3′-UCCUCC-5′。所有已知大肠杆菌 mRNA 的翻译起始区都含有 4~5 个（至少 3 个）对应于 SD 序列的核苷酸。表 3-8 是噬菌体 ΦX174 基因的 5′非编码区序列，黑体显示不同 mRNA 与 16S rRNA 3′端六核苷酸保守区的同一性。




随着另一个顺反子的翻译，mRNA 的二级结构被破坏

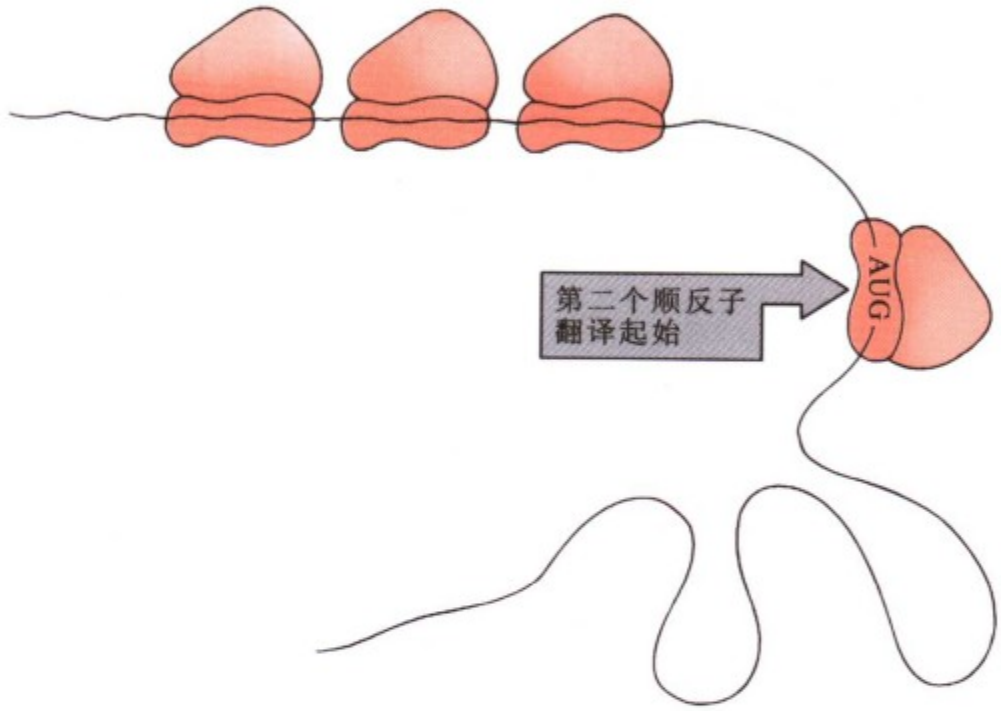



图 3-15 mRNA 的次级结构有可能控制翻译的起始

表 3-8 ϕX174 mRNA 5′端与 16S rRNA 3′端序列比较

| ϕX174 基因 | mRNA 序列（5′→3′） |
|---|---|
| D | CCACUAAU***AGG***UAAGAAAUCAUGAGU |
| E | CUGCGUUG***AGG***CUUGCGUUUAUGGUA |
| J | CGUGCGGA***AGGAG***UGAUGUAAUGUCU |
| F | CCCUUACUU***GAGG***AUAAAUUAUGUCU |
| G | UUCUGCUU***AGGAG***UUUAAUCAUGUUU |
| A | CAAAUCUU***GGAGG***CUUUUUUAUGGUU |
| B | AAAGGUCU***AGGAG***CUAAAGAAUGGAA |
| rRNA（3′→5′） | AU***UCCUCC***ACUAG |

细菌16S rRNA 的 3′端既非常保守，又高度自我互补，能形成“发卡”式结构(图 3-16)。与 mRNA 互补的部分也参与这个“发卡”的形成。表面上看这似乎是一对矛盾，一个序列不可能在形成“发卡”的同时又与 mRNA 相结合。事实上，这正好说明序列配对是动态的、可变的，翻译起始复合物的生成很可能包含了rRNA 3′端的“发卡”结构的改变，启动蛋白质翻译以后，mRNA-rRNA 杂合体被打破，为核糖体在 mRNA 模板上的移动创造了条件。

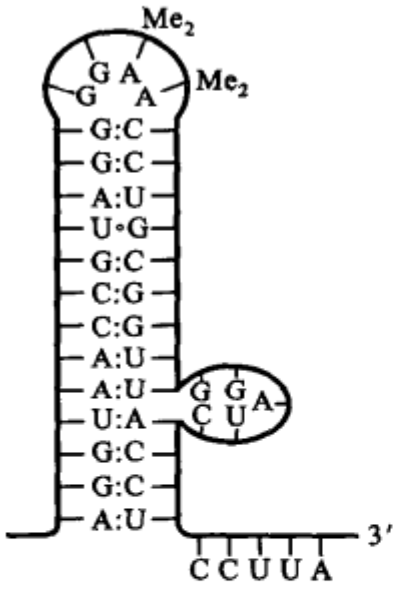



图 3-16 大肠杆菌 16S rRNA3′端的发卡式结构

## 3.4.2 真核生物 mRNA 的特征

凡是编码功能蛋白的真核基因都通过 RNA 聚合酶Ⅱ进行转录，真核基因几乎都是单顺反子，只包含一个蛋白质的信息，其长度在几百到几千个核苷酸之间。一个完整的基因，不但包括编码区(coding region)，还包括 5′和 3′端长度不等的特异性序列，它们虽然不编码氨基酸，却在基因表达的过程中起着重要作用。所以，“基因”的分子生物学定义是：产生一条多肽链或功能 RNA 所必需的全部核苷酸序列！真核生物 mRNA 结构上的最大特征是 5′端的帽子及 3′的 poly A 结构。

1. 真核生物 mRNA 的 5′端存在“帽子”结构

真核生物的 mRNA(不包括叶绿体和线粒体)5′端都是经过修饰的，基因转录一般从嘌呤(主要是 A，也可能是 G)起始，第一个核苷酸保留了 5′端的三磷酸基团并能通过其 3′-OH 位与下一个核苷酸的 5′磷酸基团形成二酯键，转录产物的起始序列为 pppApNpNp……然而，如果在体外用核酸酶处理成熟 mRNA，其 5′端并不产生预期的核苷三磷酸，而产生以 5′-5′三磷酸基团相连的二核苷酸，5′终端是一个在 mRNA 转录后加上去的甲基化鸟嘌呤(图 3-17)。

mRNA5′端加“G”的反应是由腺苷酸转移酶完成的，这个反应非常迅速，很难在体外或体内测得 5′自由三磷酸基团的存在。据测算，在新生 mRNA 链达到 50 个核苷酸之前，甚至可能在 RNA 聚合酶Ⅱ离开转录起始位点之前，帽子结构就已加到 mRNA 的第一个核苷酸上了。这就是说，mRNA 几乎一诞生就是戴上帽子的。一般认为，帽子结构是 GTP 和原 mRNA5′三磷酸腺苷(或鸟苷)缩合反应的产物，新加上的 G 与 mRNA 链上所有其他核苷酸方向正好相反，像一顶帽子倒扣在 mRNA 链上，故而得名。

mRNA 的帽子结构常常被甲基化。第一个甲基出现在所有真核细胞的 mRNA 中(单细胞真核生物 mRNA 主要是这个结构)，由鸟嘌呤-7-甲基转移酶(guanine-7-methyl-transferase)催化，称为零类帽子(cap0)。帽子结构的下一步是

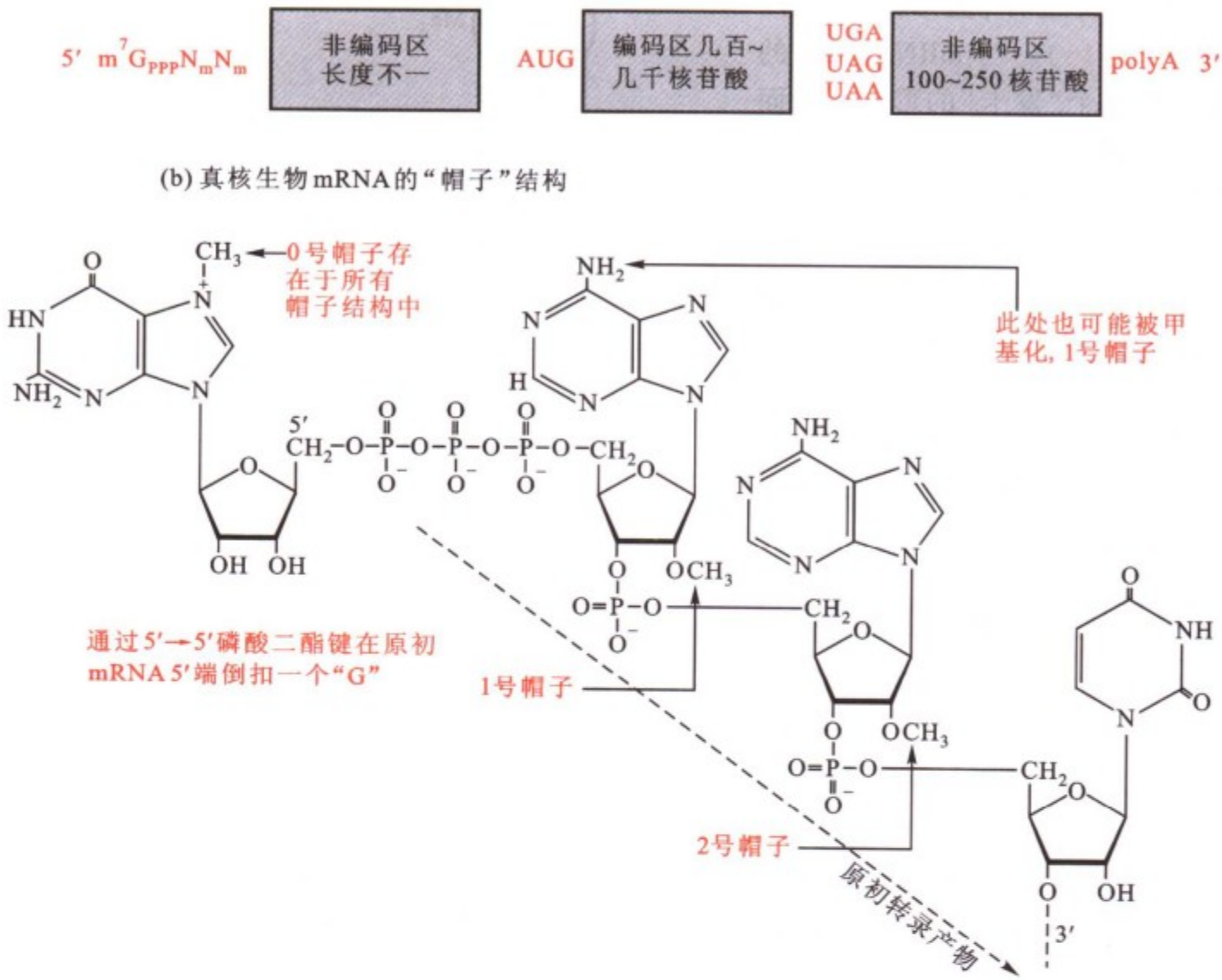



图 3-17 真核生物 mRNA 结构

在第二个核苷酸(原 mRNA 5′第一位)的 2′-OH 位上加另一个甲基,这步反应由 2′-*O*-甲基转移酶完成。一般把有这两个甲基的结构称为 1 类帽子(cap1),真核生物中以这类帽子结构为主。当 mRNA 原第二位核苷酸是腺嘌呤时,其 $N^6$ 位有时也被甲基化,这一反应只能在 2′-OH 被甲基化以后才能发生。在某些生物细胞内,mRNA 链上的第三个核苷酸的 2′-OH 位也可能被甲基化,因为这个反应只以戴有 1 类帽子的 mRNA 为底物,所以被称为 2 类帽子(cap2)。有 2 类帽子的 mRNA 只占有帽 mRNA 总量的 10%~15%以下。

帽子结构可能使 mRNA 免遭核酸酶的破坏。实验表明,去除珠蛋白 mRNA 5′端的 7-甲基鸟嘌呤后,该 mRNA 分子的翻译活性和稳定性都明显下降。而且,有帽子结构的 mRNA 更容易被蛋白质合成的起始因子所识别,从而促进蛋白质的合成。在呼肠孤病毒中,含甲基化 5′端 mRNA 的蛋白质合成速度比不含甲基的 mRNA 要快。用化学方法除去 5′端的甲基以后,上述 mRNA 作为蛋白质合成模板的活性消失,说明 mRNA 5′端甲基化的帽子是翻译所必需的。已经发现在呼肠孤病毒中无 $m^7G$ 的 mRNA 不能与核糖体 40S 小亚基结合,证明甲基化的帽子结构可能是蛋白质合成起始信号的一部分。

2. 绝大多数真核生物 mRNA 具有多聚 A 尾巴

除组蛋白基因外，真核生物 mRNA 的 3′端都有 poly A 序列，其长度因 mRNA 种类不同而变化，一般为 40~200 个左右。poly A 序列是在转录后加上去的，可能在细胞核中的不均一核 RNA 阶段就已经加上了 poly A。目前还不清楚 RNA 聚合酶Ⅱ所转录基因的精确终止位点，但研究发现，几乎所有真核基因的 3′端转录终止位点上游 15~30 bp 处的保守序列 AAUAAA 对于初级转录产物的准确切割及加 poly A 是必需的（图 3-18）。实验证明，尽管 poly A 位点及 AATAAA 的存在对于基因的转录和成熟意义重大，但 RNA 聚合酶Ⅱ却不在 poly A 位点终止，而往往继续转录，因此，大部分已知基因的初级转录产物拥有 poly A 位点下游 0.5~2 kb 核苷酸序列。

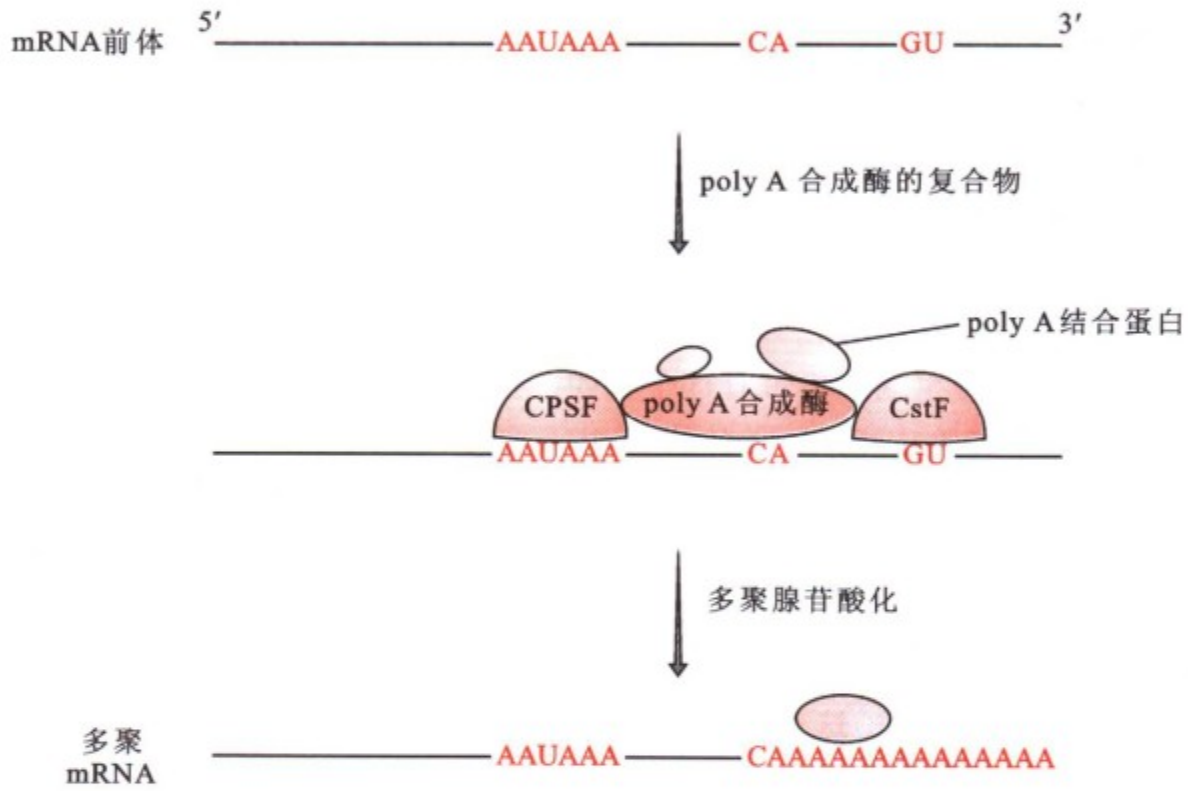



图 3-18 真核生物 mRNA 的加 poly A 反应

加 poly A 时需要由内切酶切开 mRNA 3′端的特定部位，然后由 poly A 合成酶催化多聚腺苷酸的反应。如果将上述保守序列切除，该基因组合的转录活性就会消失。点突变实验将 AAUAAA 变为 AAGAAA，虽然维持了该基因的转录活性，却发现 mRNA 的剪接加工受阻，因而没有功能性 mRNA 产生。

poly A 是 mRNA 由细胞核进入细胞质所必需的形式，它大大提高了 mRNA 在细胞质中的稳定性。当 mRNA 刚从细胞核进入细胞质时，其 poly A 尾巴一般比较长，随着 mRNA 在细胞质内逗留时间延长，poly A 逐渐变短消失，mRNA 进入降解过程。

真核生物 mRNA 大都具有 poly A 尾巴，这一特性已被广泛应用于分子克隆。常用寡聚 dT 片段与 mRNA 上的 poly A 相配对，作为反转录酶合成第一条 cDNA 链的引物。

尽管大部分真核 mRNA 有 poly A 尾巴，细胞中仍可有多达 1/3 没有 poly A

的 mRNA,把带有 poly A 的 mRNA 称为 poly $A^+$,而把没有 poly A 的 mRNA 称为 poly $A^-$。现已查明,约 1/3 的 poly $A^-$ mRNA 编码了不同形式的组蛋白,其余 2/3 的 poly $A^-$ mRNA 可能带有与 poly $A^+$组分相同的遗传信息。

## 3.5 终止和抗终止

到目前为止,科学家尚未发现某个单一位点具有特异的转录终止功能。但是,小鼠 $\beta$-珠蛋白基因的终止区(terminator)能被用来终止腺病毒基因的转录,农杆菌胭脂碱合成酶基因终止区能被用来终止几乎所有的外源基因,说明可能存在共同的转录终止机制。

一般情况下,RNA 聚合酶起始基因转录后,它就会沿着模板 5′→3′方向不停地移动,合成 RNA 链,直到碰上终止信号时才与模板 DNA 相脱离并释放新生 RNA 链。终止发生时,所有参与形成 RNA-DNA 杂合体的氢键都必须被破坏,模板 DNA 链才能与有意义链重新组合成 DNA 双链。

研究 RNA 终止时最经常遇到的问题是 3′端核苷酸的定位,因为活细胞内部根据终止信号正确终止的 RNA 与一个经过剪接的 RNA 的 3′端没有两样,都是—OH 基团,而研究 5′时,只有初生 RNA 上才有一个三磷酸基团,任何经过剪接修饰的 RNA 的 5′端不再带有这个基团。

根据体外实验中 RNA 聚合酶是否需要辅助因子参与才能终止 RNA 链的延伸,可将大肠杆菌中的终止子分为不依赖于 ρ 因子和依赖于 ρ 因子两大类。

### 3.5.1 不依赖于 ρ 因子的终止

在不依赖于 ρ 因子这类终止反应中,没有任何其他因子的参与,核心酶也能在某些位点终止转录。现已查明,模板 DNA 上存在终止转录的特殊信号——终止子,又称内在终止子(intrinsic terminator),每个基因或操纵子都有一个启动子和一个终止子。

内在终止子有两个明显的结构特点:①终止位点上游一般存在一个富含 GC 碱基的二重对称区,由这段 DNA 转录产生的 RNA 容易形成发卡式结构。②在终止位点前面有一段由 4~8 个 A 组成的序列,所以转录产物的 3′端为寡聚 U,这种结构特征的存在决定了转录的终止。在新生 RNA 中出现发卡式结构会导致 RNA 聚合酶的暂停,破坏 RNA-DNA 杂合链 5′端的正常结构。寡聚 U 的存在使杂合链的 3′端部分出现不稳定的 rU·dA 区域。两者共同作用使 RNA 从三元复合物中解离出来(图 3-19)。

终止效率与二重对称序列和寡聚 U 的长短有关,随着发卡式结构(至少 6 bp)和寡聚 U 序列(至少 4 个 U)长度的增加,终止效率逐步提高。

### 3.5.2 依赖于 ρ 因子的终止

有些终止位点的 DNA 序列缺乏共性,而且不能形成强的发卡结构,因而不

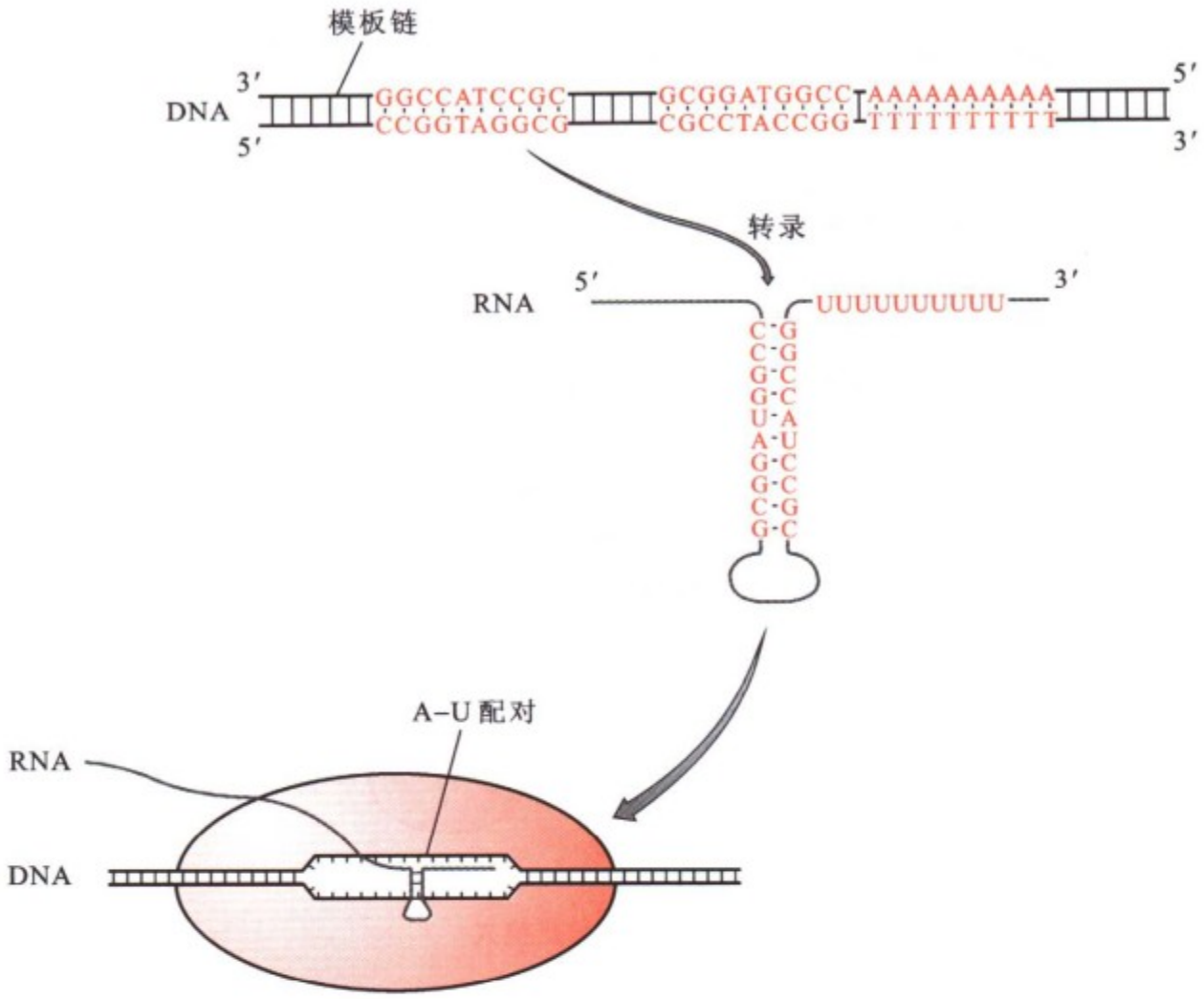



图 3–19 由基因序列决定的转录终止过程示意图

能诱导转录的自发终止。体外转录实验表明,RNA 聚合酶并不能识别这些转录终止信号，只有在加入大肠杆菌 ρ 因子后该聚合酶才能在 DNA 模板上准确地终止转录。ρ 因子是一个由 6 个相同亚基组成的六聚体,它的相对分子质量约为 275 000,它具有 NTP 酶和解螺旋酶活性,能水解各种核苷酸三磷酸,它通过催化 NTP 的水解促使新生 RNA 链从三元转录复合物中解离出来,从而终止转录。大肠杆菌的 ρ–依赖型终止子占所有终止子的一半左右。

目前认为,ρ 因子是 RNA 聚合酶终止转录的重要辅助因子，它的作用机制可用“穷追”(hot pursuit)模型来解释。RNA 合成起始以后,ρ 因子即附着在新生的 RNA 链 5′端的某个可能有序列或二级结构特异性的位点上,利用 ATP 水解产生的能量,沿着 5′→3′方向朝转录泡靠近,其运动速度可能比 RNA 聚合酶移动的速度快些;当 RNA 聚合酶移动到终止子而暂停时,ρ 因子到达 RNA 的 3′–OH 端追上并取代了暂停在终止位点上的 RNA 聚合酶,它所具有的 RNA–DNA 解螺旋活性使转录产物 RNA 从模板 DNA 上释放,随后,转录复合物解体,完成转录过程(图 3–20)。

### 3.5.3 抗终止

由于不同的生理要求,在转录过程中有时即使遇到终止信号,仍然需要继续转录,于是出现了抗转录终止现象。抗终止作用最早是在噬菌体感染细菌中

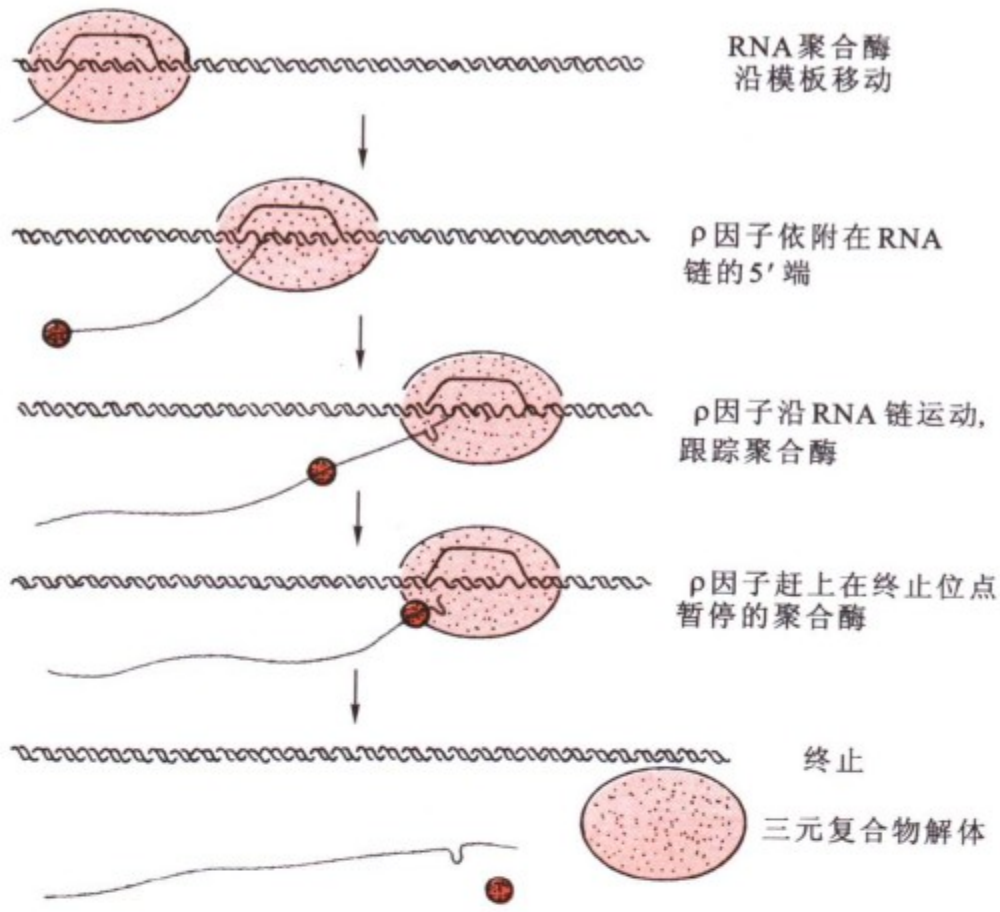



图 3-20 ρ 因子参与的 RNA 合成终止模式

发现的,它是细菌操纵子和噬菌体在调控回路中不同于上述两种方式的从另一个角度对转录进行的调控。抗转录终止主要有两种方式:

1. 破坏终止位点 RNA 的茎-环结构

在一些控制氨基酸合成的操纵子结构基因前面有一段前导序列,具有终止信号,中间有串联的编码某一氨基酸的密码子,由于转录与翻译是偶联的,转录产物 mRNA 形成后立即通过核糖体指导蛋白质合成,当介质中该氨基酸浓度较高时,与此相对应的氨酰-tRNA 含量也较高,因此,核糖体能顺利通过串联密码子。在这种情况下,mRNA 形成正常的二级结构, 其中包括末端的茎-环结构,RNA 聚合酶终止转录。当介质中该氨基酸浓度较低时,缺乏相应的氨酰-tRNA,致使核糖体滞留在串联密码子上,mRNA 不能形成特定的二级结构,末端的茎-环结构被破坏,因此转录仍能进行下去,出现转录的抗终止现象。

2. 依赖于蛋白质因子的转录抗终止

λ 噬菌体中由 *N* 基因编码产生的 N 蛋白具有抗转录终止作用,但是其功能的发挥依赖于宿主所产生的 NusA,NusB,S10 和 $2.5\times10^4$ 等几种蛋白质,这些蛋白质结合到终止子附近的 DNA 位点是实现抗转录终止的必要条件。该 DNA 位点中含有 A 区(CGCTCTTA)和一个二重对称序列,N 蛋白能识别后者转录所形成的茎-环结构并与之结合。NusA 以二聚体的形式存在,其一个亚基与 RNA 聚合酶结合,另一个亚基与 N 蛋白结合,当其他 3 种蛋白都结合到 Nut 位点时,便形成一个蛋白复合物,并通过 NusA 与 RNA 聚合酶结合,改变聚合酶的构象,使之对终止信号不敏感,继续催化 RNA 链的合成(图 3-21)。

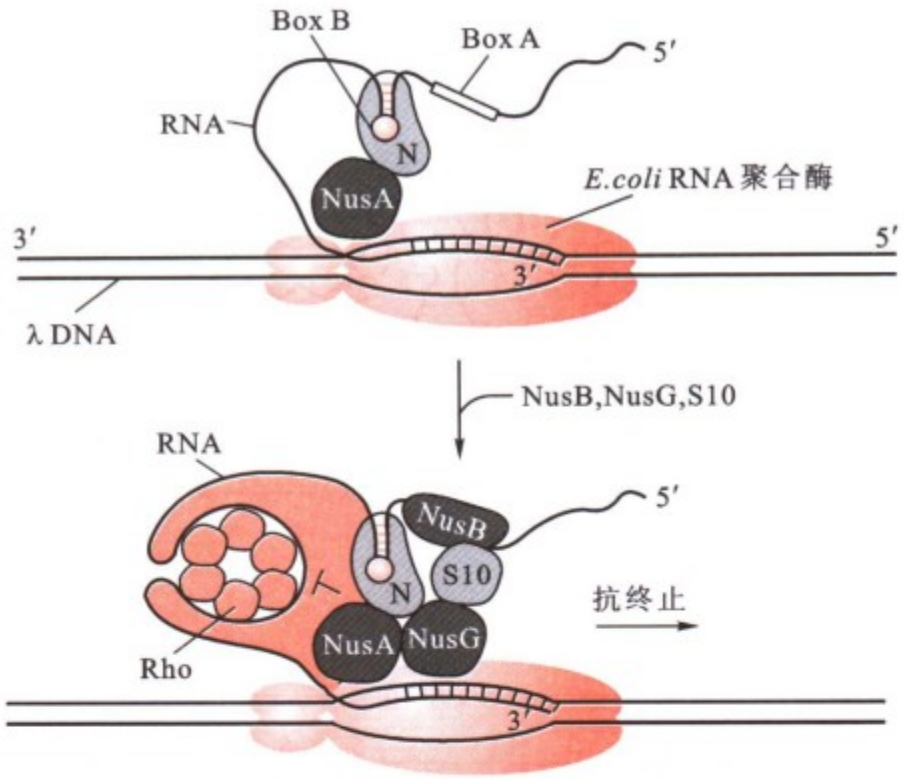



图 3-21 依赖于蛋白质因子的转录抗终止模式

## 3.6 内含子的剪接、编辑、再编码及化学修饰

### 3.6.1 RNA 中的内含子

断裂基因的存在表明真核细胞的基因结构和 mRNA 合成过程比原核细胞要复杂得多，因为真核基因表达往往伴随着 RNA 的剪接过程（splicing），从 mRNA 前体分子中切除被称为内含子(intron)的非编码区，并使基因中被称为外显子(exon)的编码区拼接形成成熟 mRNA(图 3-22)。

真核基因大多是断裂的，也就是说，一个基因可由多个内含子和外显子间隔排列而成。研究表明，内含子在真核基因中所占的比例很高，甚至超过 99%(表 3-9)。

由DNA 转录生成的原始转录产物——核不均一 RNA(heterogeneous nuclear RNA，hnRNA)，即 mRNA 的前体，经过 5′加“帽”和 3′酶切加多聚腺苷酸，再经过 RNA 的剪接，编码蛋白质的外显子部分就连接成为一个连续的可译框架(open reading frame，ORF)，通过核孔进入细胞质，就能作为蛋白质合成的模板了。表 3-10 和图 3-23 分别总结了原始 RNA 成熟期间的主要加工过程。

真核基因平均含有 8~10 个内含子，前体分子一般比成熟 mRNA 大 4~10 倍。不同生物细胞内含子的边界处存在相似的核苷酸序列，表明内含子剪接过程在进化上是保守的。比较同源基因的进化过程发现内含子的异化大于外显子，特定的内含子还可能在进化过程中丢失，因此，内含子的“功能”及其在生物进化中的地位是一个引人注目的问题。另外，许多人类疾病是内含子剪接异常引起的，如地中海贫血患者的珠蛋白基因中，大约有 1/4 的核苷酸突变发生在内

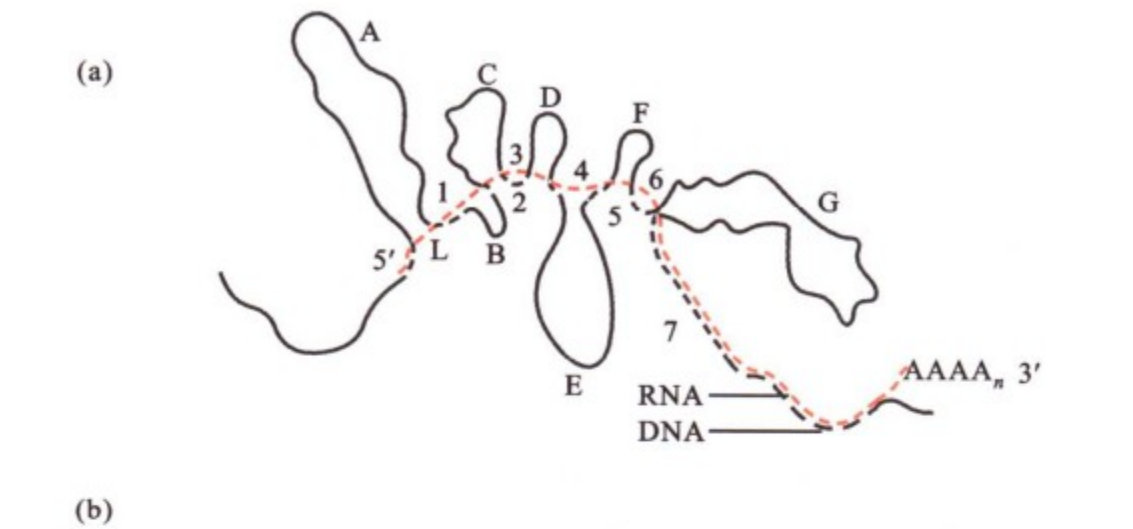



图 3-22 用DNA-RNA 杂交的方法验证鸡卵清蛋白基因中存在非编码的内含子区 (a) 成熟 mRNA 与带有该基因的互补链 DNA 序列杂交示意图;(b) 鸡卵清蛋白的基因结构示意图,L 及 1~7 区表示外显子区,A~G 为内含子区。

表 3-9 部分人类基因中内含子序列所占的比例分析

| 基 因 | 长度/kb | 内含子数量 | 内含子所占比例/% |
|---|---|---|---|
| 胰岛素 | 1.4 | 2 | 67 |
| β-球蛋白 | 1.4 | 2 | 69 |
| 血清蛋白 | 18 | 13 | 89 |
| 胶原蛋白组分Ⅶ | 31 | 117 | 71 |
| Ⅷ因子 | 186 | 25 | 95 |
| 萎缩性肌强直因子 | 2 400 | 78 | >99 |

表 3-10 RNA 加工过程及其生理功能

| 加工过程 | 推测的生理功能 |
|---|---|
| 加帽子反应 | mRNA 从细胞核向细胞质运转,翻译起始 |
| 加 polyA 反应 | 转录终止,翻译起始和 mRNA 降解 |
| RNA 的剪接 | 从 mRNA,tRNA 和 rRNA 分子中切除内含子 |
| RNA 的切割 | 从前体 RNA 中释放成熟 tRNA 和 rRNA 分子 |

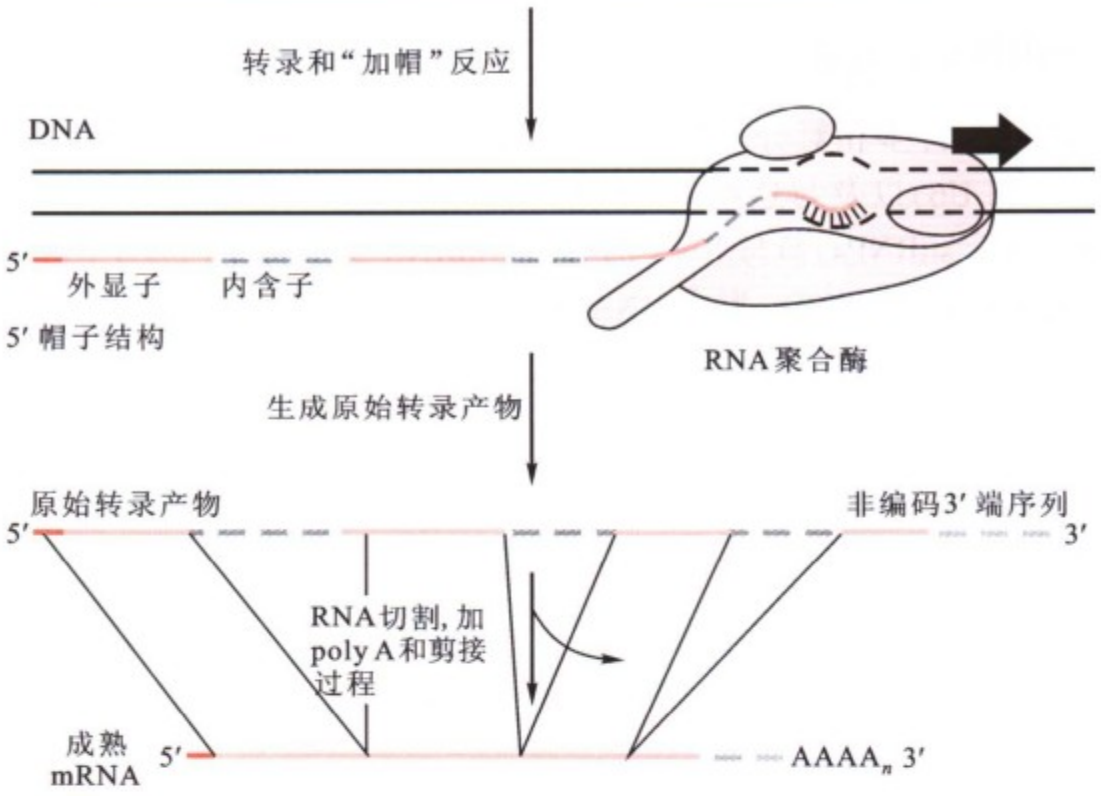



图 3-23 原始转录产物的生成及其主要加工剪接过程图示

含子的 5′或 3′边界保守序列上,或者虽然位于内含子中间但干扰了前体 mRNA 的正常剪接。

比较不同基因的核苷酸序列发现,mRNA 前体中内含子的两端边界存在共同的序列，这些序列结构可能是产生 mRNA 前体剪接的信号。多数细胞核 mRNA 前体中内含子的 5′边界序列为 GU,3′边界序列为 AG。因此,GU 表示供体衔接点的 5′端,AG 代表接纳体衔接点的 3′端。把这种保守序列模式称为 GU-AG 法则,又称为 Chambon 法则。

除了边界序列之外,外显子与内含子交界处的序列,内含子内部的部分序列也可能参与内含子的剪接(图 3-24)。在 3′端剪接位点 AG 的附近有一段富含嘧啶的区域,含有 10~20 个嘧啶核苷酸;5′端有一保守序列(5′-GUPuAGU-3′)。此外,许多发生分叉剪接的核 mRNA 内含子 3′端上游 18~50 个核苷酸处,存在一个序列为 $Py_{80}NPy_{87}Pu_{75}APy_{95}$ 的保守区，其中 A 为百分之百保守，且具有 2′-OH，是参与形成分叉剪接中间物的特定腺嘌呤。上述保守序列都是pre-mRNA 剪接过程中各种核糖核蛋白剪接调节因子的结合位点,对于有效核准确的剪接非常重要。

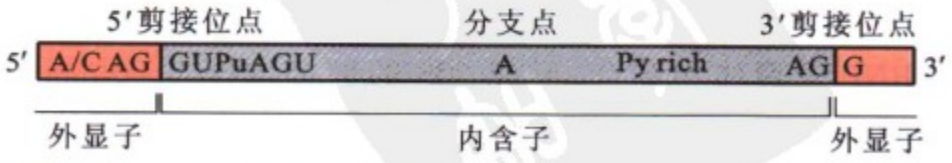



图 3-24 脊椎动物 mRNA 前体中常见的内含子剪接所必需的保守序列 图中显示 5′剪接位点和 3′剪接位点的共有序列以及分支点的保守腺苷酸成分。

### 3.6.2 mRNA 的剪接

研究表明,许多相对分子质量较小(长约 106~185bp)的核内 RNA(如 U1,U2,U4,U5 和 U6)以及与这些 RNA 相结合的核蛋白(small nuclear ribonucleoprotein particle,snRNPs)参与 RNA 的剪接。mRNA 链上每个内含子的 5′和 3′端分别与不同的 snRNP 结合,形成 RNA 和 RNP 复合物(图 3-25)。

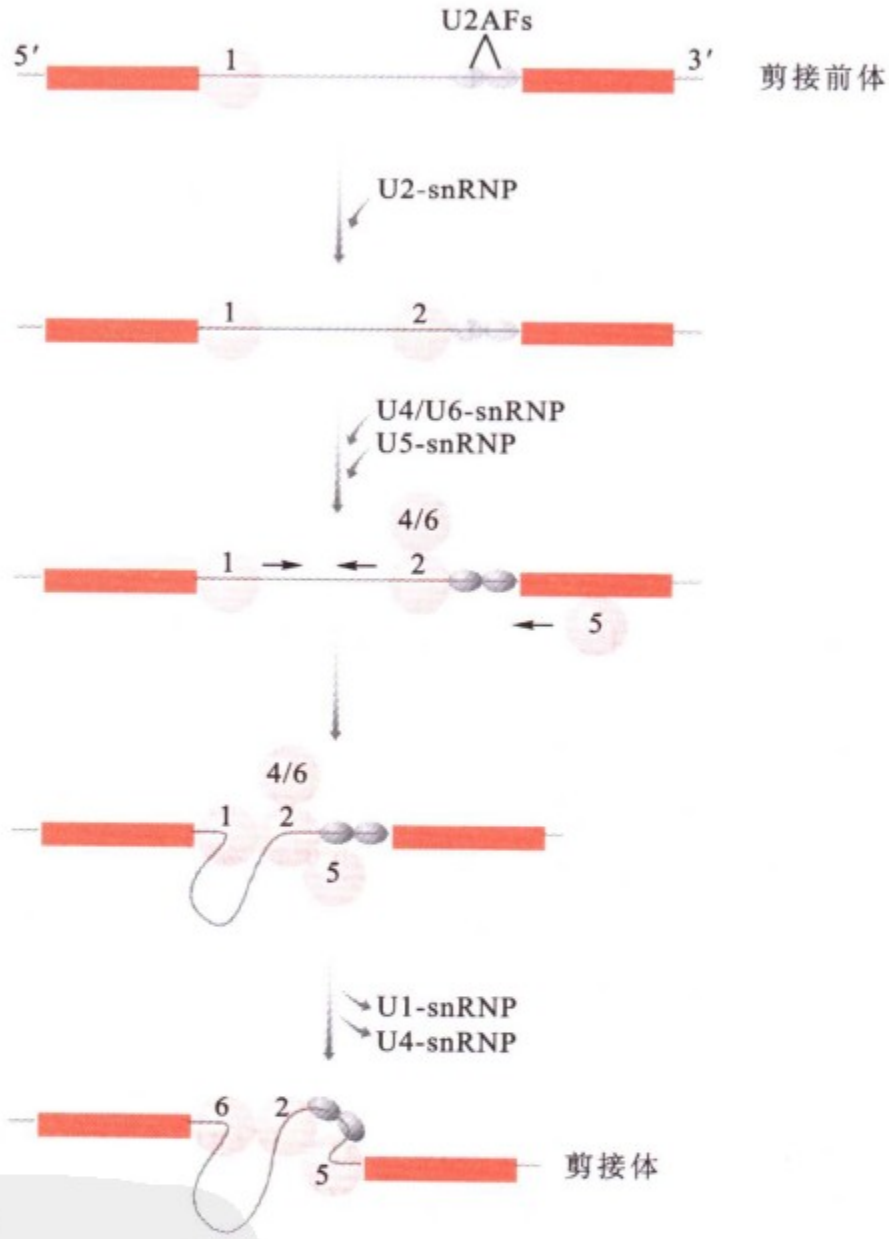



图 3-25 真核生物 mRNA 前体中内含子剪接过程示意图

一般情况下,由 U1 snoRNA 以碱基互补的方式识别 mRNA 前体 5′剪接点,由结合在 3′剪接点上游富嘧啶区的 U2AF(U2 auxiliary factor)识别 3′剪接点并引导 U2 snRNP 与分支点相结合,形成剪接前体(pre-spliceosome),并进一步与 U4、U5、U6 snRNP 三聚体相结合,形成 60S 的剪接体(spliceosome),进行 RNA 前体分子的剪接。哺乳动物细胞中,mRNA 前体上的 snRNP 是从 5′向下游“扫描”,选择在分支点富嘧啶区 3′下游的第一个 AG 作为剪接的 3′位点。AG 前一位核苷酸可以影响剪接效率,一般说来,CAG=UAG>AAG>GAG。如果 mRNA 前体上同时存在几个 AG,可能发生剪接竞争。

在高等真核生物中，内含子通常是有序或组成性地从 mRNA 前体中被剪接，然而，在个体发育或细胞分化时可以有选择性地越过某些外显子或某个剪接点进行变位剪接，产生出组织或发育阶段特异性 mRNA，称为内含子的变位剪接。图 3-26 是果蝇中发现的与性别分化相关基因的特异性剪接。雌果蝇 *sxl* 基因前体 mRNA 中的第 3 个外显子直接被切除，形成功能型 SXL 蛋白(外显子

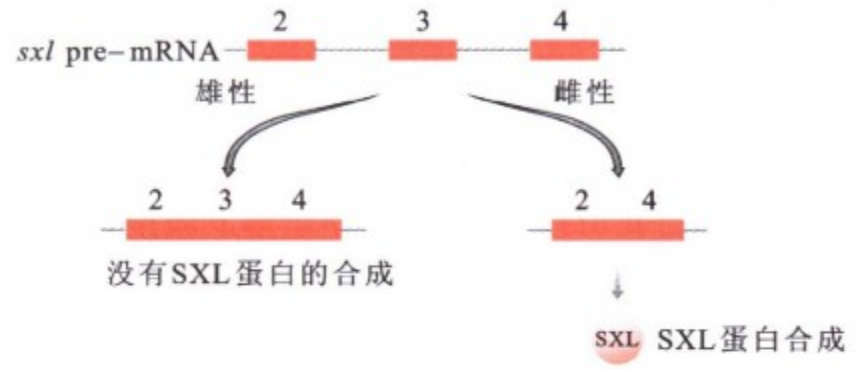



**图 3-26 果蝇中与性别分化相关基因产物的特异性剪接过程** (a) *sxl* mRNA 前体在不同性别的果蝇中发生选择性剪接；(b) SXL 蛋白使 $U2AF^{65}$ 与雌果蝇 *tra* 基因第 2 个外显子中的隐性切割位点相结合，发生选择性剪接，产生功能型 TRA 蛋白；(c) TRA 蛋白进一步诱导产生了性状特异的 DSX蛋白。

2 和 4 相连)。由于这个蛋白的存在,使雌果蝇 *tra* 基因第二个外显子的切割位点发生变化,产生功能型 TRA 蛋白。最后,在功能型 TRA 蛋白和一组 SR 蛋白(富含丝氨酸和精氨酸蛋白)的作用下,特异性剪接 *dsx* 基因的前体 mRNA,产生雄性和雌性特征性 DSX 蛋白,参与性别分化中的一系列生理生物化学反应。脊椎动物中大约有 5%的基因能以这种方式进行剪接,保证各同源蛋白质之间既具有大致相同的结构或功能域,又具有特定的性质差异,这无疑大大拓展了基因所携带的遗传信息。

与前面所述的 mRNA 前体中主要(GU–AG 类)和次要(AU–AC 类)内含子的剪接方式不同的是Ⅰ、Ⅱ类内含子,因为带有这些内含子的 RNA 本身具有催化活性,能进行内含子的自我剪接,而无需借助于形成剪接体。表 3–11 总结了存在于生物体内的各种内含子。

表 3–11 生物体内的各种内含子

| 内含子类型 | 细胞内定位 |
|---|---|
| GU – AG | 细胞核,pre-mRNA (真核) |
| AU – AC | 细胞核,pre-mRNA(真核) |
| Ⅰ类内含子 | 细胞核,pre-rRNA(真核),细胞器 RNA,少数细菌 RNA |
| Ⅱ类内含子 | 细胞器 RNA,部分细菌 RNA |
| Ⅲ类内含子 | 细胞器 RNA |
| 双内含子 | 细胞器 RNA |
| pre – tRNA 中的内含子 | 细胞核,pre – tRNA(真核) |

最初在研究原生动物四膜虫 pre-RNA 时发现了Ⅰ类内含子,后来在细菌中也发现存在这类内含子。Ⅰ类内含子的剪接主要是转酯反应(trans-esterification),即剪接反应实际上是发生了两次磷酸二酯键的转移。在Ⅰ类内含子切除体系中,第一个转酯反应由一个游离的鸟苷或鸟苷酸(GMP、GDP 或 GTP)介导,鸟苷或鸟苷酸的 3′–OH 作为亲核基团攻击内含子 5′端的磷酸二酯键,从上游切开 RNA 链。在第二个转酯反应中,上游外显子的自由 3′–OH 作为亲核基团攻击内含子 3′位核苷酸上的磷酸二酯键,使内含子被完全切开,上下游两个外显子通过新的磷酸二酯键相连(图 3–27)。

Ⅱ类内含子主要存在于真核生物的线粒体和叶绿体 *rRNA* 基因中。在Ⅱ类内含子切除体系中,转酯反应无需游离鸟苷酸或鸟苷,而是由内含子本身的靠近 3′端的腺苷酸 2′–OH 作为亲核基团攻击内含子 5′端的磷酸二酯键,从上游切开 RNA 链后形成套索状结构。再由上游外显子的自由 3′–OH 作为亲核基团攻击内含子 3′位核苷酸上的磷酸二酯键,使内含子被完全切开,上下游两个外显子通过新的磷酸二酯键相连(图 3–28)。

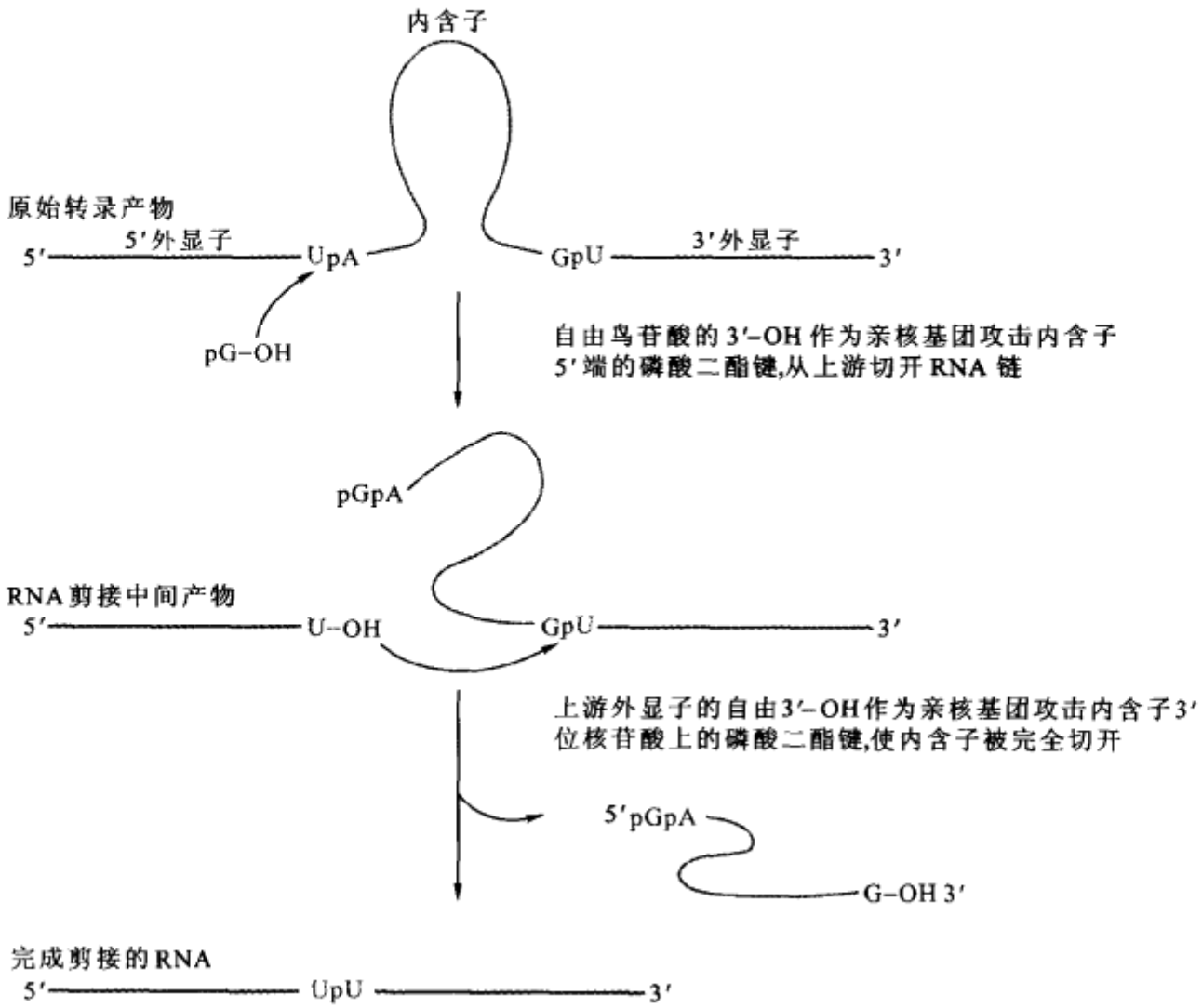



图 3-27 I 类内含子的自我剪接过程

## 3.6.3 RNA 的编辑、再编码及化学修饰

1. RNA 的编辑

根据中心法则,DNA、RNA 和蛋白质之间存在着直接的线性关系,即连续的序列 DNA 被真实地转录成 mRNA 序列,然后翻译产生蛋白质分子。断裂基因的发现和 RNA 的剪接虽然使得基因表达过程增加了一个步骤,但 DNA 的实际编码序列没有发生变化。

RNA 的编辑(RNA editing)是某些 RNA,特别是 mRNA 前体的一种加工方式,如插入、删除或取代一些核苷酸残基,导致 DNA 所编码的遗传信息的改变,因为经过编辑的 mRNA 序列发生了不同于模板 DNA 的变化。介导 RNA 编辑的机制有两种:位点特异性脱氨基作用和引导 RNA 指导的尿嘧啶插入或删除。

哺乳动物载脂蛋白 mRNA 的编辑是广泛研究的典型例子,图 3-29 分别是该基因的 DNA 序列以及在哺乳动物肝脏和肠组织中分离到的 mRNA 序列。载脂蛋白基因编码区共有 4 563 个密码子,在所有组织中的 DNA 序列都相同。在肝脏中,该基因转录产生完整的 mRNA 并被翻译成有 4 563 个氨基酸的全长蛋白质,相对分子质量为 $5.1\times10^5$。在肠中合成的却是只包含有 2 153 个密码子的

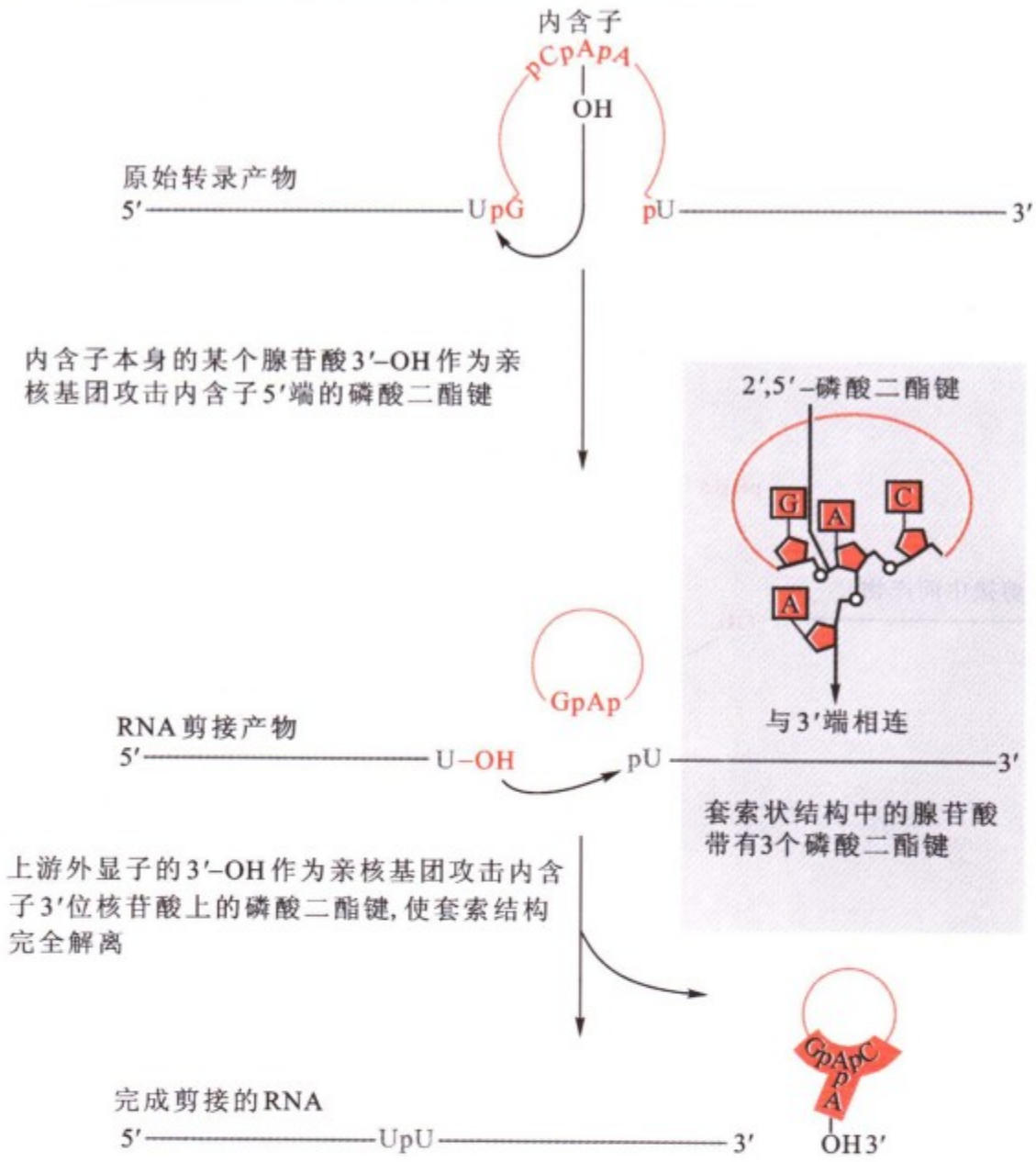



图 3–28 II 类内含子的自我剪接过程

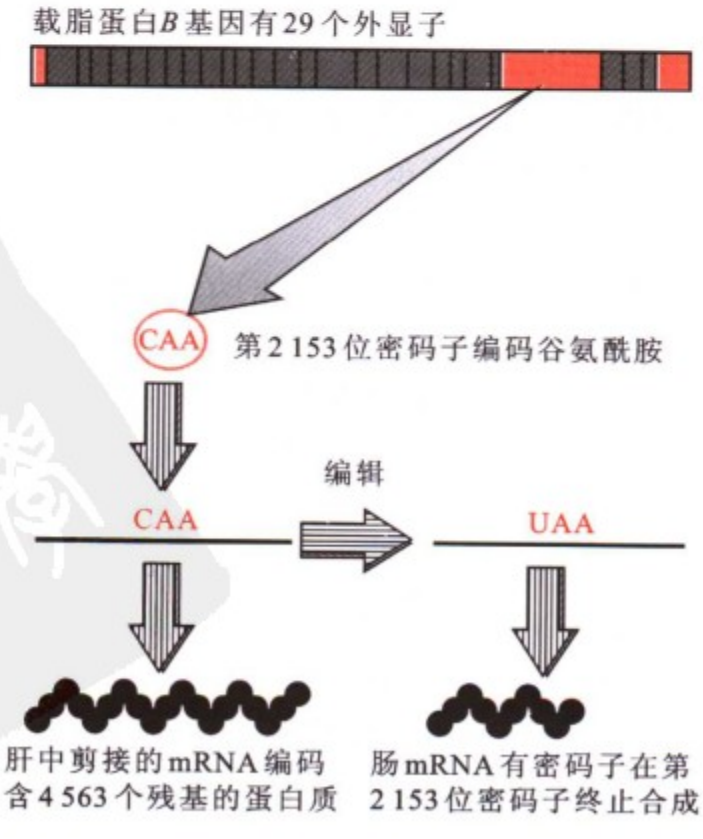



图 3–29 哺乳动物载脂蛋白基因转录产物的编辑

mRNA，翻译产生相对分子质量为 $2.5\times10^5$ 的蛋白质。研究发现，该蛋白其实是全长载脂蛋白的 N 端，它是由一个在序列上除了 2 153 位密码子从 CAA 突变为 UAA 之外完全与肝脏 mRNA 相同的核酸分子所编码的，C→U 突变使编码谷氨酰胺的密码子变成了终止密码子。

RNA 的编辑虽然不是很普遍，在真核生物中也时有发生(表 3-12)。大鼠脑中谷氨酸受体蛋白 mRNA 经编辑后，分子中有多个编码谷氨酸的密码子变成了在控制通过神经递质的离子流过程中有重要影响的精氨酸，表明 RNA 的编辑可能是充分发挥生理功能所必需的。

表 3-12 哺乳动物中 RNA 编辑的实例

| 组织 | 靶标 RNA | 所改变的碱基 | 结　果 |
|---|---|---|---|
| 肝脏，肠 | 载脂蛋白 B | C→U | 谷氨酰胺密码子→终止子 |
| 肌肉 | 半乳糖苷酶 | U→A | 苯丙氨酸密码子→酪氨酸 |
| 睾丸，肿瘤等 | Wilms 肿瘤基因 -1 | U→C | 亮氨酸密码子→脯氨酸 |
| 肿瘤 | 神经纤维瘤基因 -1 | C→U | 精氨酸密码子→终止子 |
| 脑 | 谷氨酸受体蛋白 | A→I | 多个谷氨酸密码子→精氨酸 |

载脂蛋白 mRNA 中的 C→U，谷氨酸受体蛋白 mRNA 中的 A→I，都属于脱氨基作用的结果，分别由胞嘧啶和腺嘌呤脱氨酶所催化。通常情况下，该酶促反应的特异性不强，腺嘌呤脱氨酶可以作用于双链 RNA 区的任何腺苷酸残基。但是，RNA 的编辑发生在带有具催化作用的脱氨酶亚基的复合体中，有附加的 RNA 结合区能帮助识别所编辑的特异性靶位点。

除单碱基突变之外，RNA 编辑的另一种形式是尿苷酸的缺失和添加。研究发现，利什曼原虫属细胞色素 b mRNA 中含有许多独立于核基因的尿嘧啶残基，而特异性插入这些残基的信息来自指导 RNA(guide RNA)，因为它含有与编辑后细胞色素 b mRNA 相互补的核苷酸序列。指导 RNA 与被编辑区及其周围部分核酸序列虽然有相当程度的互补性，但该 RNA 上存在一些未能配对的腺嘌呤，形成缺口，为插入尿嘧啶提供了模板(图 3-30)。反应完成后，指导 RNA 从 mRNA 上解离下来，而 mRNA 则被用做翻译的模板。

RNA 编辑具有重要的生物学意义：

(1) 校正作用　有些基因在突变过程中丢失的遗传信息可能通过 RNA 的编辑得以恢复。

(2) 调控翻译　通过编辑可以构建或去除起始密码子和终止密码子，是基因表达调控的一种方式。

(3) 扩充遗传信息　能使基因产物获得新的结构核功能，有利于生物的进化。

2. RNA 的再编码

有研究发现，mRNA 在某些情况下不是以固定的方式被翻译，而可以改变原来的编码信息，以不同的方式进行翻译，科学上把 RNA 编码和读码方式的改变称为 RNA 的再编码(RNA recoding)。

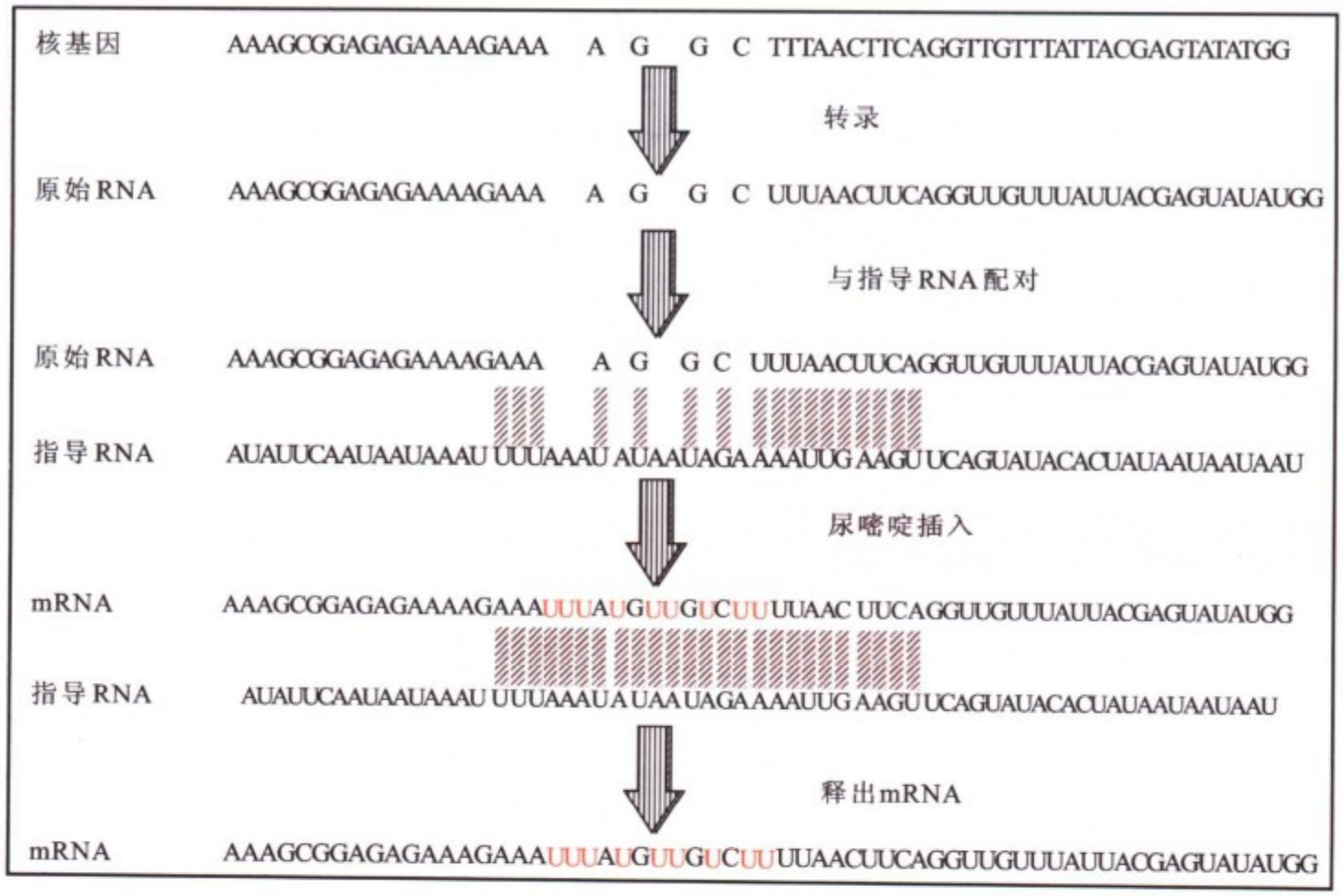



图 3–30 指导 RNA 和 RNA 的编辑机制

某些有机体中进行蛋白质翻译时,mRNA 的读码信号在 tRNA、rRNA 和其他蛋白因子作用下发生+1/–1 移位，甚至发生核糖体跳过 50 个核苷酸的情况，前两者称为核糖体程序性+1/–1 移位,后者称为核糖体跳跃。最近人们还发现自然中存在第二十一种氨基酸硒代半胱氨酸(selenocysteine)和第二十二种氨基酸吡咯赖氨酸(pyrrolysine),两者都是通过终止子通读而编码的。上述的+1/–1 移码、核糖体跳跃以及终止子通读都是再编码的表现方式,因为它们不同于常规基因的解码规则。RNA 的再编码可以从一个 mRNA 产生两种或多种相互关联但又不同的蛋白质,这也可能是蛋白质合成的一种调节机制。

3. RNA 的化学修饰

除了 RNA 的编辑之外,有些 RNA,特别是前体 rRNA 和 tRNA,还可能有特异性化学修饰。图 3–31 是最常见的 6 大类化学修饰途径及其产物。研究证实,仅人细胞内 rRNA 分子上就存在 106 种甲基化和 95 种假尿嘧啶产物，虽然尚未发现特征性保守序列,但有实验证据表明,RNA 的化学修饰可能具有位点特异性。有实验表明,只含有 70~100 个核苷酸的核仁 RNA(snoRNAs)参与 RNA 的化学修饰,因为这些 RNA 能通过碱基配对的方式,把 rRNA 分子上需要修饰的位点找出来(图 3–32)。现在一般认为,snoRNA 上的 D 盒是甲基化酶的识别位点。

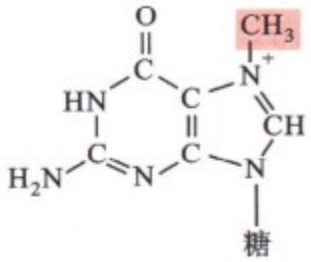
### 3.6.4 核酶

核酶（ribozyme）是指一类具有催化功能的 RNA 分子，通过催化靶位点

**甲基化**

在核苷酸的碱基或核糖基上加一个或多个$-CH_3$

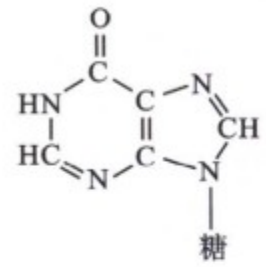
举例：鸟嘌呤甲基化产物 $m^7G$




**去氨基化**

从碱基上去掉氨基

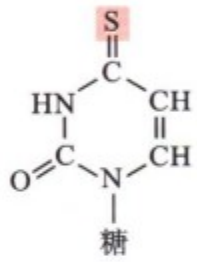
举例：鸟嘌呤去氨基后成为次黄嘌呤




**硫代**

用硫取代碱基分子上的氧

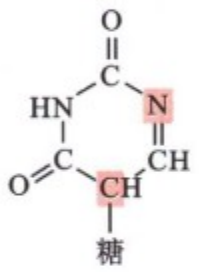
举例：4-硫尿嘧啶




**碱基的同分异构化**

碱基环结构上发生分子替代

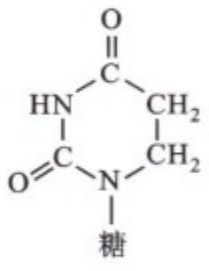
举例：尿嘧啶变构生成假尿嘧啶




**二价键的饱和化**

把一个二价键饱和

举例：二氢尿嘧啶




**核苷酸的替代**

用不常见核苷酸替换常见核苷酸

举例：奎嘌呤

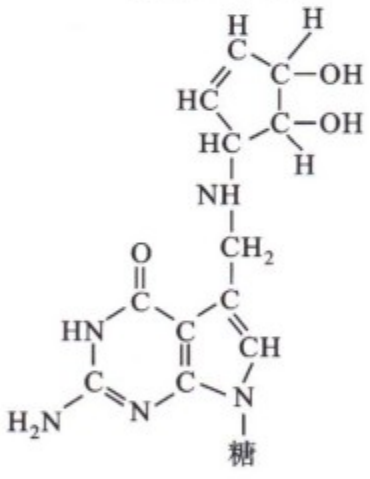



图 3-31 rRNA 和 tRNA 分子中核苷酸的化学修饰

图 3-32 酵母 U24 snoRNA 指导的甲基化

RNA链中磷酸二酯键的断裂,特异性地剪切底物 RNA 分子,从而阻断基因的表达。ribozyme 是 ribonucleic acid(核糖核酸)和 enzyme(酶)两个词的缩合词。

1982 年 Cech 从四膜虫 rRNA 前体的加工研究中首先发现 rRNA 前体具有自我剪接作用,于是提出 ribozyme 的概念,打破了酶都是蛋白质这种传统的观念。1983 年,Altman 等人发现 RNaseP 中的 RNA 组分可以催化 tRNA 前体的加工,之后的研究还发现 T4 RNA、一些植物的类病毒(viroid)、拟病毒(virusoid)等都能进行自我剪接。

核酶的催化功能与其空间结构有密切关系,目前已知有多种特殊结构的核酶,如 RNaseP 的 RNA 亚基(M1 RNA)、锤头型、发夹型丁型肝炎 δ 病毒 RNA、Ⅰ类内含子和Ⅱ类内含子等。具有自我剪切能力的 RNA 大多数都能形成锤头结构(hammerhead structure)。该二级结构由三个茎(Ⅰ、Ⅱ、Ⅲ)构成,茎区是由互补碱基构成的局部双链结构,包围着一个由 11~13 个保守核苷酸构成的催化中心(图 3-33)。核酶剪切通常发生在 H 位点。同一核酶分子由具有催化中心的核酶和含有剪切位点的底物部分共同组成锤头结构,底物部分是切割部位两端的核苷酸,它与核酶的茎Ⅰ和茎Ⅲ结合,在切割之后该底物被释放,由一个新的没有被切割的底物取代,使切割反应得以重复进行。

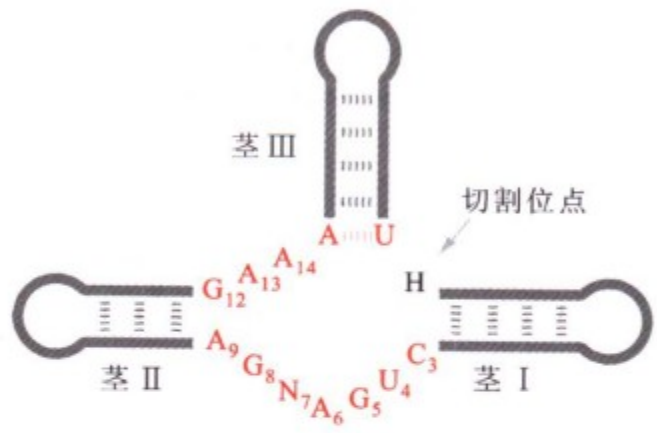



**图 3-33 核酶的锤头结构** 该酶包括三个茎环区,其间有一个 11~13 个保守核苷酸构成的催化中心。N代表该位置上可以为任意碱基,箭头代表切割部位,H 代表除了 G 以外的核苷酸。在类病毒中,茎Ⅲ的顶部并没有被环连接起来。

不同的核酶具有不同的催化功能,可以分为剪切型核酶和剪接型核酶两大类。

1. 剪切型核酶

剪切型核酶是只剪不接,能够催化自身 RNA 或不同的 RNA 分子,切下特异的核苷酸序列。M1 RNA 是最典型的例子,在高浓度 $Mg^{2+}$存在时,它可特异性地剪切 tRNA 前体 5′端片段。此外,四膜虫 rRNA 前体的剪接产物 L19 也是一种剪切型核酶。

2. 剪接型核酶

该类核酶具有序列特异的内切核酸酶、RNA 连接酶等多种酶的活性,它既能切割 RNA 分子,也能通过转酯反应形成新的磷酸二酯键,连接切割后的 RNA 分子。前面介绍的Ⅰ类和Ⅱ类内含子就属于剪接型核酶。

锤头型核酶和发夹型核酶都能够催化产物的连接,但发夹型核酶催化连接反应的活性比催化切割反应的活性高约 10 倍,而锤头型核酶催化切割反应的活性比催化连接反应的活性高 100 倍。核酶的发现为基因治疗提供了新的策

略，因为根据其自我剪切的特点，可以人工合成多种核酶以抑制破坏病毒基因或癌基因等有害基因的功能。

核酶的发现使我们对 RNA 的重要功能又有了新的认识。因此，核酶是继反转录现象之后对中心法则的又一个重要修正，说明 RNA 既是遗传物质又是酶。核酶的发现为生命起源的研究提供了新思路——也许曾经存在以 RNA 为基础的原始生命。照这么说，蛋白质世界也可能(仅仅是可能)起源于 RNA 世界！

### 3.6.5 RNA 在生物进化中的地位

我们曾在本章开篇时说过，基因是唯一能自主复制、世代相传并永久存在的遗传学单位。中心法则也阐明了 DNA 作为遗传信息的贮存者，是生命遗传和生物进化的“中心”。那么，RNA(或许还包括蛋白质！)有没有可能在生物遗传变异的历史演变中起作用呢？

研究发现，RNA 不仅具有极大的拷贝数，其变异率高达 $10^{-4}$~$10^{-3}$ 核苷酸/RNA 分子，翻译错误率累加可达 $10^{-2}$~$10^{-1}$ 氨基酸/蛋白质分子，能够产生各种不同蛋白质而迅速积累信息。这就是说，RNA 在基因表达过程中有一定的离散，正如散弹枪射击会有一些铅弹偏离靶心一样。RNA 比相应 DNA 序列含有更多的遗传信息，可通过剪接、改变和校正可译框架等方式表达出多种蛋白质异构体(同工酶？)，还可能通过反转录产生与 RNA 信息相一致的 DNA 分子，直接影响后代的基因型，这种作用也可称为“反馈效应”。

其次，RNA 还是获得性遗传的分子基础。所谓获得性遗传是指有机体在生长发育过程中由于环境的影响而不是由于基因突变所形成的新的遗传性状。在3类主要生物大分子中，DNA 是信息分子，蛋白质是功能分子，RNA 则既是信息分子，又是功能分子，它在表达过程中起着信息提取和加工的作用。环境可以影响和诱导 RNA 的产生和信息加工，从而使机体表现出新的性状，这种性状如果传递给子代细胞和个体，就产生获得性遗传。实际上，不仅核酸分子是遗传信息的载体，某些亚细胞结构、细胞器和生物膜也能传递部分遗传信息，它们在环境影响下发生的结构改变有时也能传递给后代。

## 思考题

1. 什么是编码链？什么是模板链？
2. 简述 RNA 转录的概念及其基本过程。
3. 大肠杆菌的 RNA 聚合酶有哪些组成成分？各个亚基的作用如何？
4. 什么是封闭复合物、开放复合物以及三元复合物？
5. 简述 σ 因子的作用。
6. 什么是 Pribnow box？它的保守序列是什么？
7. 什么是上升突变？什么是下降突变？
8. 简述原核生物和真核生物 mRNA 的区别。
9. 大肠杆菌的终止子有哪两大类？请分别介绍一下它们的结构特点。

10. 真核生物的原始转录产物必须经过哪些加工才能成为成熟 mRNA,以用作蛋白质合成的模板?
11. 简述Ⅰ、Ⅱ类内含子的剪接特点。
12. 什么是 RNA 编辑? 其生物学意义是什么?
13. 核酶具有哪些结构特点? 其生物学意义是什么?

## 主要参考文献


1. Aphasizhev R,Sbicego S,Peris M. *et al.* Trypanosome mitochondrial 3′terminal uridylyl transferase(TUTase):the key enzyme in U-insertion/deletion RNA editing. Cell,2002,108:637–648.
2. Bar-Nahum G,Nudler E. Isolation and characterization of sigma(70)-retaining transcription elongation complexes from Escherichia coli. Cell,2001,106:443–451.
3. Blackwood E M,Kadonaga J T. Going the distance:a current view of enhancer action. Science,1998,281:60–63.
4. Cech T R. Self-splicing RNA:implications for evolution. Int. Rev. Cytol. 1985,93:3–22.
5. Doherty E A,Doudna J A. Ribozyme Structures and Mechanisms.Ann. Rev. Biochem. 2000,69:597–615.
6. Greenblatt J,Nodwell J R,Mason S W. Transcriptional antitermination. Nature,1993,364:401–406.
7. Helmann J D,Chamberlin M. Structure and function of bacterial sigma factors. Annu. Rev. Biochem. 1988,57:839–872.
8. Korzheva N,Mustaev A,Kozlov M,*et al.* A structural model of transcription elongation. Science,2000,289:619–625.
9. Krämer A. The structure and function of proteins involved in mammalian pre-mRNA splicing.Annu. Rev. Biochem. 1996,65:367–409.
10. Sharp P A. Splicing of mRNA precursors. Science,1987,235:766–771.
11. von Hippel P H. An Integrated Model of the Transcription Complex in Elongation,Termination,and Editing Science,1998,281:660–665.

# 4 生物信息的传递(下)

## ——从 mRNA 到蛋白质


- 4.1 遗传密码——三联子
- 4.2 tRNA
- 4.3 核糖体
- 4.4 蛋白质合成的生物学机制
- 4.5 蛋白质转运机制


蛋白质是基因表达的最终产物,它的生物合成是一个比 DNA 复制和转录更为复杂的过程,主要包括:

① 翻译的起始——核糖体与 mRNA 结合并与氨基酰-tRNA 生成起始复合物。

② 肽链的延伸——由于核糖体沿 mRNA 5′端向 3′端移动,开始了从 N 端向 C 端的多肽合成,这是蛋白质合成过程中速度最快的阶段。

③ 肽链的终止及释放。核糖体从 mRNA 上解离,准备新一轮合成反应。

核糖体是蛋白质合成的场所,mRNA 是蛋白质合成的模板,转移 RNA(transfer RNA,tRNA)是模板与氨基酸之间的接合体。此外,在合成的各个阶段还有许多蛋白质、酶和其他生物大分子参与。例如,在真核生物细胞中有 70 种以上的核糖体蛋白质,20 种以上的氨酰-tRNA 合成酶,10 多种起始因子、延伸因子及终止因子,50 种左右的 tRNA 及各种rRNA、mRNA和 100 种以上翻译后加工酶参与蛋白质合成和加工过程。蛋白质合成是一个需能反应,要有各种高能化合物的参与。据统计,在真核生物中有将近 300 种生物大分子与蛋白质的生物合成有关,细胞所用来进行合成代谢总能量的 90%消耗在蛋白质合成过程中,而参与蛋白质合成的各种组分约占细胞干质量的 35%。

在真核生物细胞核内合成的 mRNA,只有被运送到细胞质部分才能被翻译生成蛋白质。所谓翻译是指将 mRNA 链上的核苷酸从一个特定的起始位点开始,按每 3 个核苷酸代表一个氨基酸的

原则,依次合成一条多肽链的过程。尽管蛋白质合成过程十分复杂,但合成速度却高得惊人,如大肠杆菌只需要 5 s 就能合成一条由 100 个氨基酸组成的多肽,而且每个细胞中成百上千个蛋白质的合成都是有条不紊地协同进行的。

## 4.1 遗传密码——三联子

贮存在 DNA 上的遗传信息通过 mRNA 传递到蛋白质上,mRNA 与蛋白质之间的联系是通过遗传密码的破译来实现的。mRNA 上每 3 个核苷酸翻译成蛋白质多肽链上的一个氨基酸,这 3 个核苷酸就称为密码,也叫三联子密码(codon)。翻译时从起始密码子 AUG(initiation codon)开始,沿着 mRNA5′→3′的方向连续阅读密码子,直至终止密码子(termination codon)为止,生成一条具有特定序列的多肽链——蛋白质(图 4-1)。这条由起始密码子开始,到终止密码子结束的核酸序列就被称为一个可译框架(open reading frame)或可读框。新生的多肽链中氨基酸的组成和排列顺序决定于其 DNA(基因)的碱基组成及其顺序。因此,作为基因产物的蛋白质最终是受基因控制的。

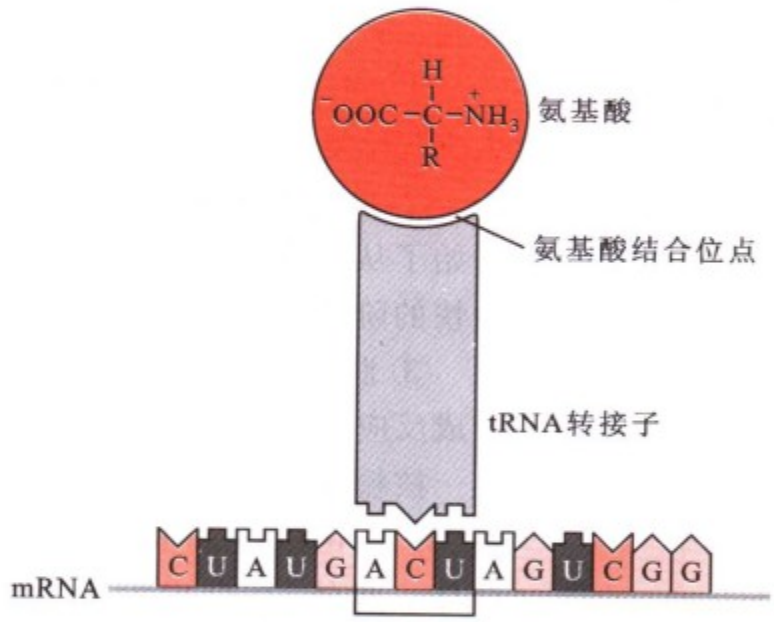



图 4-1 Crick 关于 tRNA 分子破译 mRNA 遗传密码三联子的原始构想

遗传密码是 20 世纪 60 年代通过设计出色的生物化学和遗传学实验阐明的,它是科学史上的杰出成就之一,不仅为研究蛋白质的生物合成提供了理论依据,也证实了中心法则的正确性。70 年代以来,分子生物学技术如 DNA、RNA 序列测定及氨基酸序列测定技术的进步,使遗传密码的存在得到验证。下面我们介绍遗传密码的来源及其特性。

### 4.1.1 三联子密码及其破译

蛋白质中的氨基酸序列是由 mRNA 中的核苷酸序列决定的,所以,要

知道它们之间的关系就要弄清核苷酸和氨基酸数目的对应关系。mRNA 中只有 4 种核苷酸，而蛋白质中有 20 种氨基酸，以一种核苷酸代表一种氨基酸是不可能的。若以两种核苷酸作为一个氨基酸的密码(二联子)，它们能代表的氨基酸有 $4^2=16$ 种，还是不足 20 种。而假定以 3 个核苷酸代表一个氨基酸，则可以有 $4^3=64$ 种密码，完全可以满足编码 20 种氨基酸的需要。

经过反复研究，Crick 等人首先从遗传学的角度证实三联子密码的构想是正确的。他们发现 T4 噬菌体 rⅡ位点上两个基因的正确表达与它能否侵染大肠杆菌有关，用吖啶类试剂(诱导核苷酸插入或从 DNA 链上丢失)处理使 T4 噬菌体 DNA 发生移码突变(frameshift mutation)，从而指导生成一个完全不同的、没有功能的蛋白质，使噬菌体丧失感染能力。若在模板 mRNA 中插入或删除一个碱基，会改变该密码子以后的全部氨基酸序列。若同时对模板进行插入和删除试验，保证后续密码子序列不变，翻译得到的蛋白质序列就保持不变(除了发生突变的那个密码子所代表的氨基酸之外)。如果同时删去 3 个核苷酸，翻译产生少了一个氨基酸的蛋白质，但序列不发生变化(图 4-2)。另外，对烟草坏死卫星病毒的研究发现，其外壳蛋白亚基由 400 个氨基酸组成，而相应的 RNA 片段长约 1 200 个核苷酸，与假设的密码三联子体系正好相吻合。

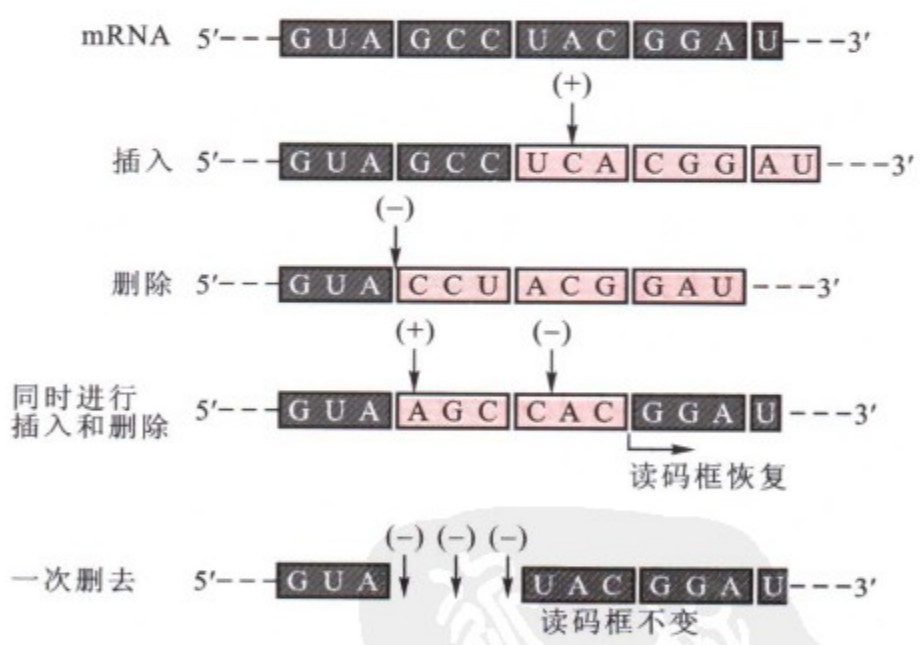



**图 4-2 用核苷酸的插入或删除实验证明 mRNA 模板上每 3 个核苷酸组成一个密码子**

遗传密码的破译，即确定代表每种氨基酸的具体密码，在 20 世纪 60 年代初期是一项困难的任务。尽管如此，由于体外蛋白质合成体系的建立和核酸人工合成技术的发展，科学家实际上只花了几年时间就解开了这个谜。

1. 以均聚物、随机共聚物和特定序列的共聚物为模板指导多肽的合成

制备大肠杆菌的无细胞合成体系，在含 DNA、mRNA、tRNA、核糖体、AA-

tRNA 合成酶及其他酶类的抽提物中加入 DNase，降解体系中的 DNA，耗尽 mRNA 时，体系中的蛋白质合成即停止，当补充外源 mRNA 或人工合成的各种均聚物或共聚物作为模板以及 ATP、GTP、氨基酸等成分时又能合成新的肽链，新生肽链的氨基酸顺序由外加的模板所决定。因此，分析比较加入的模板和合成的肽链即可推知编码某些氨基酸的密码。

1961 年，Nirenberg 把 poly U 作为模板加入到上述无细胞体系时意外地发现，新合成的多肽链是多聚苯丙氨酸，从而认定 UUU 代表苯丙氨酸(Phe)。以 poly C 及 poly A 做模板得到的分别是多聚脯氨酸和多聚赖氨酸，这样很快就解决了 3 个氨基酸的密码子(当然，一个氨基酸可能有几个密码子)。幸运的是，由于上述无细胞体系中 $Mg^{2+}$浓度很高，人工合成的多聚核苷酸不需要起始密码子就能指导多肽的生物合成，读码起始是随机的。在生理 $Mg^{2+}$条件下，没有起始密码子的多核苷酸不能被用作多肽合成的模板。

Nirenberg 及 Ochoa 等又用各种随机的共聚物或特定序列共聚物作模板合成多肽，例如，以只含 A、C 的共聚核苷酸作模板，任意排列时可出现 8 种三联子，即 CCC、CCA、CAC、ACC、CAA、ACA、AAC、AAA，获得由 Asn、His、Pro、Gln、Thr、Lys 等 6 种氨基酸组成的多肽。

他们发现，以多聚二核苷酸作模板可合成由 2 个氨基酸组成的多肽，如以poly UG 为模板合成的是多聚 Cys 和 Val，因为 poly UG 中含 Cys 和 Val 的密码：

5′…UGU GUG UGU GUG UGU GUG…3′

无论读码从 U 开始还是从 G 开始，都只能有 UGU(Cys)及 GUG(Val)两种密码子。

以多聚三核苷酸作为模板可得到有 3 种氨基酸组成的多肽。如以 poly UUC 为模板，可能有 3 种起读方式：

5′…UUC UUC UUC UUC UUC…3′

或 5′…UCU UCU UCU UCU UCU…3′

或 5′…CUU CUU CUU CUU CUU…3′

根据读码起点不同，产生的密码子可能是 UUC(Phe)、UCU(Ser)或 CUU(Leu)，所以得到的多肽可能是多聚苯丙氨酸、多聚丝氨酸或多聚亮氨酸，由此可知 UUC、UCU、CUU 分别是苯丙氨、丝氨酸及亮氨酸的密码子。当然，以多聚三核苷酸为模板时也可能只合成 2 种均聚多肽，以 poly GUA 为例：

5′…GUA GUA GUA GUA GUA…3′

或 5′…UAG UAG UAG UAG UAG…3′

或 5′…AGU AGU AGU AGU AGU…3′

由第二种读码方式产生的密码子 UAG 是终止密码，不编码任何氨基酸，因此，只产生 2 种密码子 GUA(Val)或 AGU(Ser)，所以合成的多肽要么是多聚缬氨酸，要么是多聚丝氨酸。

2. 核糖体结合技术

Nirenberg 和 Leder 还用核糖体结合技术来解决密码问题。这个方法是以人

工合成的三核苷酸如 UUU、UCU、UGU 等为模板，在含核糖体、AA-tRNA 的适当离子强度的反应液中保温，然后使反应液通过硝酸纤维素滤膜。他们发现，游离的 AA-tRNA 因相对分子质量小而能自由通过滤膜，加入三核苷酸模板可以促使其对应的 AA-tRNA 结合到核糖体上，体积超过膜上的微孔而被滞留，这样就能把已结合到核糖体上的 AA-tRNA 与未结合的 AA-tRNA 分开。若用 20 种 AA-tRNA 做20 组同样的实验，每组都含 20 种 AA-tRNA 和各种三核苷酸，但只有一种氨基酸用 $^{14}C$ 标记，看哪一种 AA-tRNA 被留在滤膜上，进一步分析这一组的模板是哪个三核苷酸，从模板三核苷酸与氨基酸的关系可测知该氨基酸的密码子。例如，模板是 UUU 时，Phe-tRNA 结合于核糖体上，可知 UUU 是 Phe 的密码子(表 4-1)。

表 4-1 三核苷酸密码子能使特定的氨酰 - tRNA 结合到核糖体上

| 密码子 | 与核糖体相结合的 $^{14}C$ 标记的氨酰 - tRNA | | |
|---|---|---|---|
| | Phe - $tRNA^{Phe}$ | Lys - $tRNA^{Lys}$ | Pro - $tRNA^{Pro}$ |
| UUU | 4.6 * | 0 | 0 |
| AAA | 0 | 7.7 | 0 |
| CCC | 0 | 0 | 3.1 |

* 数字代表特定氨酰 - tRNA 与带有模板三核苷酸的核糖体相结合的效率。随机结合 = 1。

## 4.1.2 遗传密码的性质

1. 密码的连续性

翻译由 mRNA 的 5′端的起始密码子开始，一个密码子接一个密码子连续地阅读直到 3′终止密码，密码间无间断也没有重叠，即起始密码子决定了所有后续密码子的位置，说明三联子密码是连续的。

2. 密码的简并性(degeneracy)

按照 1 个密码子由 3 个核苷酸组成的原则，4 种核苷酸可组成 64 个密码子，现在已经知道其中 61 个是编码氨基酸的密码子，另外 3 个即 UAA、UGA 和 UAG 并不代表任何氨基酸，它们是终止密码子，不能与 tRNA 的反密码子配对，但能被终止因子或释放因子识别，终止肽链的合成(表 4-2)。其中 UAA 叫赭石(ochre)密码，UAG 叫琥珀(amber)密码，UGA 叫蛋白石(opal)密码。细菌中三种终止密码子的使用频率是不同的，UAA 是最常用的，UGA 比 UAG 的使用频率更高一点，但 UGA 出错的可能性更大一些。

因为存在 61 种密码子而只有 20 种氨基酸，所以许多氨基酸有多个密码子，实际上除甲硫氨酸(ATG)和色氨酸(UGG)只有一个密码子外，其他氨基酸都有一个以上的密码子，其中 9 种氨基酸有 2 个密码子，1 种氨基酸有 3 个密码子，5 种氨基酸有 4 个密码子，3 种氨基酸有 6 个密码子(表 4-3)。

由一种以上密码子编码同一个氨基酸的现象称为简并，对应于同一氨基酸的密码子称为同义密码子(synonymous codon)。另外，AUG 和 GUG 既是甲硫氨

表 4-2 通用遗传密码及相应的氨基酸

| 第一位(5′端)核苷酸 | 第二位(中间)核苷酸 U | 第二位(中间)核苷酸 C | 第二位(中间)核苷酸 A | 第二位(中间)核苷酸 G | 第三位(3′端)核苷酸 |
|---|---|---|---|---|---|
| U | 苯丙氨酸 (Phe,F) | 丝氨酸 (Ser,S) | 酪氨酸 (Tyr,Y) | 半胱氨酸 (Cys,C) | U |
| | 苯丙氨酸 (Phe,F) | 丝氨酸 (Ser,S) | 酪氨酸 (Tyr,Y) | 半胱氨酸 (Cys,C) | C |
| | 亮氨酸 (Leu,L) | 丝氨酸 (Ser,S) | 终止 (Stop) | 终止 (Stop) | A |
| | 亮氨酸 (Leu,L) | 丝氨酸 (Ser,S) | 终止 (Stop) | 色氨酸 (Trp,W) | G |
| C | 亮氨酸 (Leu,L) | 脯氨酸 (Pro,P) | 组氨酸 (His,H) | 精氨酸 (Arg,R) | U |
| | 亮氨酸 (Leu,L) | 脯氨酸 (Pro,P) | 组氨酸 (His,H) | 精氨酸 (Arg,R) | C |
| | 亮氨酸 (Leu,L) | 脯氨酸 (Pro,P) | 谷氨酰胺 (Gln,Q) | 精氨酸 (Arg,R) | A |
| | 亮氨酸 (Leu,L) | 脯氨酸 (Pro,P) | 谷氨酰胺 (Gln,Q) | 精氨酸 (Arg,R) | G |
| A | 异亮氨酸 (Ile,I) | 苏氨酸 (Thr,T) | 天冬酰胺 (Asn,N) | 丝氨酸 (Ser,S) | U |
| | 异亮氨酸 (Ile,I) | 苏氨酸 (Thr,T) | 天冬酰胺 (Asn,N) | 丝氨酸 (Ser,S) | C |
| | 异亮氨酸 (Ile,I) | 苏氨酸 (Thr,T) | 赖氨酸 (Lys,K) | 精氨酸 (Arg,R) | A |
| | 甲硫氨酸 (Met,M) | 苏氨酸 (Thr,T) | 赖氨酸 (Lys,K) | 精氨酸 (Arg,R) | G |
| G | 缬氨酸 (Val,V) | 丙氨酸 (Ala,A) | 天冬氨酸 (AsP,D) | 甘氨酸 (Gly,G) | U |
| | 缬氨酸 (Val,V) | 丙氨酸 (Ala,A) | 天冬氨酸 (AsP,D) | 甘氨酸 (Gly,G) | C |
| | 缬氨酸 (Val,V) | 丙氨酸 (Ala,A) | 谷氨酸 (Glu,E) | 甘氨酸 (Gly,G) | A |
| | 缬氨酸 (Val,V) | 丙氨酸 (Ala,A) | 谷氨酸 (Glu,E) | 甘氨酸 (Gly,G) | G |

表 4-3 密码子的兼并性

| 氨基酸 | 密码子个数 | 氨基酸 | 密码子个数 |
|---|---|---|---|
| 丙氨酸 | 4 | 亮氨酸 | 6 |
| 精氨酸 | 6 | 赖氨酸 | 2 |
| 天冬酰胺 | 2 | 甲硫氨酸 | 1 |
| 天冬氨酸 | 2 | 苯丙氨酸 | 2 |
| 半胱氨酸 | 2 | 脯氨酸 | 4 |
| 谷氨酰胺 | 2 | 丝氨酸 | 6 |
| 谷氨酸 | 2 | 苏氨酸 | 4 |
| 甘氨酸 | 4 | 色氨酸 | 1 |
| 组氨酸 | 2 | 酪氨酸 | 2 |
| 异亮氨酸 | 3 | 缬氨酸 | 4 |

酸及缬氨酸的密码子又是起始密码子，这种双重功能在生物学上的意义尚不清楚。

同义密码子一般都不是随机分布的，因为其第一、第二位核苷酸往往是相同的，而第三位核苷酸的改变并不一定影响所编码的氨基酸，这种安排减少了变异对生物的影响。一般说来，编码某一氨基酸的密码子越多，该氨基酸在蛋白质中出现的频率也越高，只有精氨酸是个例外，因为在真核生物中 CG 双联子出现的频率较低，所以尽管有 4 个同义密码子，蛋白质中精氨酸的出现频率仍不高，如图 4-3。

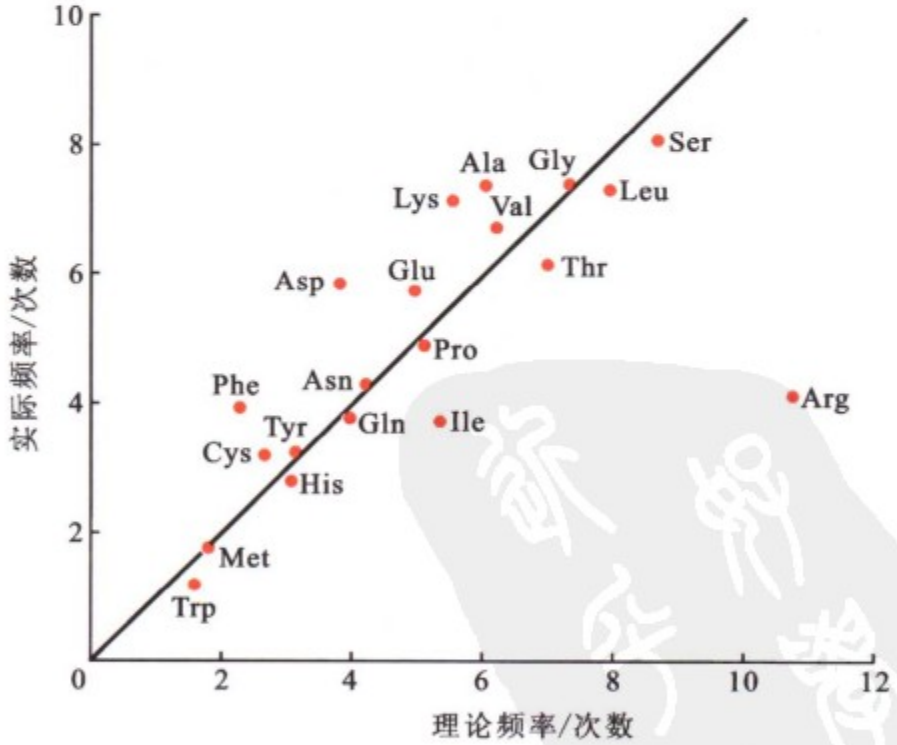



图 4-3 除了 Arg 以外，编码某一特定氨基酸的密码子个数与该氨基酸在蛋白质中的出现频率相吻合

3. 密码的通用性与特殊性

遗传密码无论在体内还是体外，也无论是对病毒、细菌、动物还是植物而言

都是通用的,所以,密码子具有通用性。20 世纪 70 年代以后对各种生物基因组的大规模测序结果也充分证明生物界基本共用同一套遗传密码。密码子的通用性有助于我们研究生物的进化。同时,遗传密码的通用性在遗传工程中得到充分运用,比如在细菌中大量表达人类的外源蛋白——胰岛素等。

虽然密码子具有通用性,但也发现极少数例外。在支原体中,终止密码子 UGA 被用来编码色氨酸;在嗜热四膜虫中,另一个终止密码子 UAA 被用来编码谷氨酰胺。对人、牛及酵母线粒体 DNA 序列和结构的研究还发现,在线粒体中也有一些例外情况(表 4-4),如 UGA 编码色氨酸,而非终止密码子;甲硫氨酸可由 AUA 编码等,体现了遗传密码的特殊性。

表 4-4 线粒体与核 DNA 密码子使用情况的比较

| 生物 | 密码子 | 线粒体 DNA 编码的氨基酸 | 核 DNA 编码的氨基酸 |
|---|---|---|---|
| 所有 | UGA | 色氨酸 | 终止子 |
| 酵母 | CUA | 苏氨酸 | 亮氨酸 |
| 果蝇 | AGA | 丝氨酸 | 精氨酸 |
| 哺乳动物 | AGA/G | 终止子 | 精氨酸 |
| 哺乳动物 | AUA | 甲硫氨酸 | 异亮氨酸 |

4. 密码子与反密码子的相互作用

在蛋白质生物合成过程中,tRNA 的反密码子在核糖体内是通过碱基的反向配对与 mRNA 上的密码子相互作用的(图 4-4)。1966 年,Crick 根据立体化学

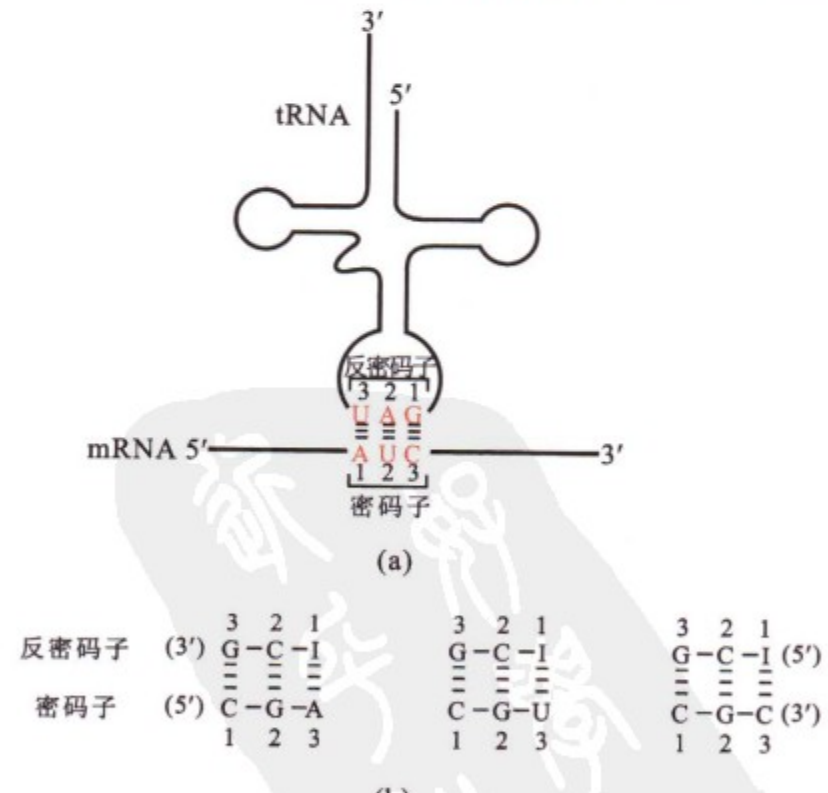



图 4-4 mRNA 上的密码子与 tRNA 上的反密码子配对示意图 (a) 密码子与 tRNA 反密码子臂上相应序列配对;(b) 当反密码子第一位是I时,密码子第三位可以是 A、U 或 C。

原理提出摆动假说(wobble hypothesis),解释了反密码子中某些稀有成分(如 I)的配对,以及许多氨基酸有 2 个以上密码子的问题。

根据摆动假说,在密码子与反密码子的配对中,前两对严格遵守碱基配对原则,第三对碱基有一定的自由度,可以"摆动",因而使某些 tRNA 可以识别 1 个以上的密码子。一个 tRNA 究竟能识别多少个密码子是由反密码子的第一位碱基的性质决定的,反密码子第一位为 A 或 C 时只能识别 1 种密码子,为 G 或 U 时可以识别 2 种密码子,为 I 时可识别 3 种密码子(表 4-5)。如果有几个密码子同时编码一个氨基酸,凡是第一、二位碱基不同的密码子都对应于各自独立的 tRNA。原核生物中大约有 30~45 种 tRNA,真核细胞中可能存在 50 种 tRNA。

表 4-5 tRNA 上的反密码子与 mRNA 上密码子的配对与"摆动"分析

| | | |
|---|---|---|
| 反密码子第一位是 C 或 A 时,只能识别一种密码子 | | |
| 反密码子 | (3′)X-Y-C(5′) | (3′)X-Y-A(5′) |
| 密码子 | (5′)Y-X-G(3′) | (5′)Y-X-U(3′) |
| 反密码子第一位是 U 或 G 时,可分别识别两种密码子 | | |
| 反密码子 | (3′)X-Y-U (5′) | (3′)X-Y-G (5′) |
| 密码子 | (5′)Y-X-A/G(3′) | (5′)Y-X-C/U(3′) |
| 反密码子第一位是 I 时,可识别 3 种密码子 | | |
| 反密码子 | (3′)X-Y-I (5′) | |
| 密码子 | (5′)Y-X-A/U/C(3′) | |

## 4.2 tRNA

tRNA 在蛋白质合成中处于关键地位,它不但为每个三联密码子翻译成氨基酸提供了接合体,还为准确无误地将所需氨基酸运送到核糖体上提供了运送载体,所以,它又被称为第二遗传密码。虽然 tRNA 分子各自的序列不同,但所有的 tRNA 都具有共同的特征:存在经过特殊的修饰碱基,tRNA 的 3′端都以 CCA-OH 结束,该位点是 tRNA 与相应氨基酸结合的位点。

tRNA 参与多种反应,并与多种蛋白质和核酸相互识别,这就决定了它们在结构上存在大量的共性。由于小片段碱基互补配对,三叶草形 tRNA 分子上有 4 条根据它们的结构或已知功能命名的手臂:

① 受体臂(acceptor arm) 主要由链两端序列碱基配对形成的杆状结构和 3′端未配对的 3~4 个碱基所组成,其 3′端的最后 3 个碱基序列永远是 CCA,最后一个碱基的 3′或 2′自由羟基(—OH)可以被氨酰化。

其余手臂均由碱基配对产生的杆状结构和无法配对的套索状结构所组成。

② $T_\psi C$ 臂是根据 3 个核苷酸命名的,其中 ψ 表示拟尿嘧啶,是 tRNA 分子所拥有的不常见核苷酸。

③ 反密码子臂是根据位于套索中央的三联反密码子命名的。

④ D 臂是根据它含有二氢尿嘧啶(dihydrouracil)命名的。

图 4-5 显示了最常见 tRNA 分子的三叶草二级结构形式。tRNA 一般有 76 个碱基,相对分子质量约为 $2.5\times10^4$。不同的 tRNA 分子可有 74~95 个核苷酸不等。tRNA 分子长度的不同主要是由其中的两条手臂引起的,在 D 臂中存在多至 3 个可变核苷酸位点,包括 17∶1(位于第 17 和 18 核苷酸之间)及 20∶1、20∶2(位于第 20 和 21 核苷酸之间)。最常见的 D 臂缺失这 3 个核苷酸,而最小的 D 臂中第 17 位核苷酸也缺失了。

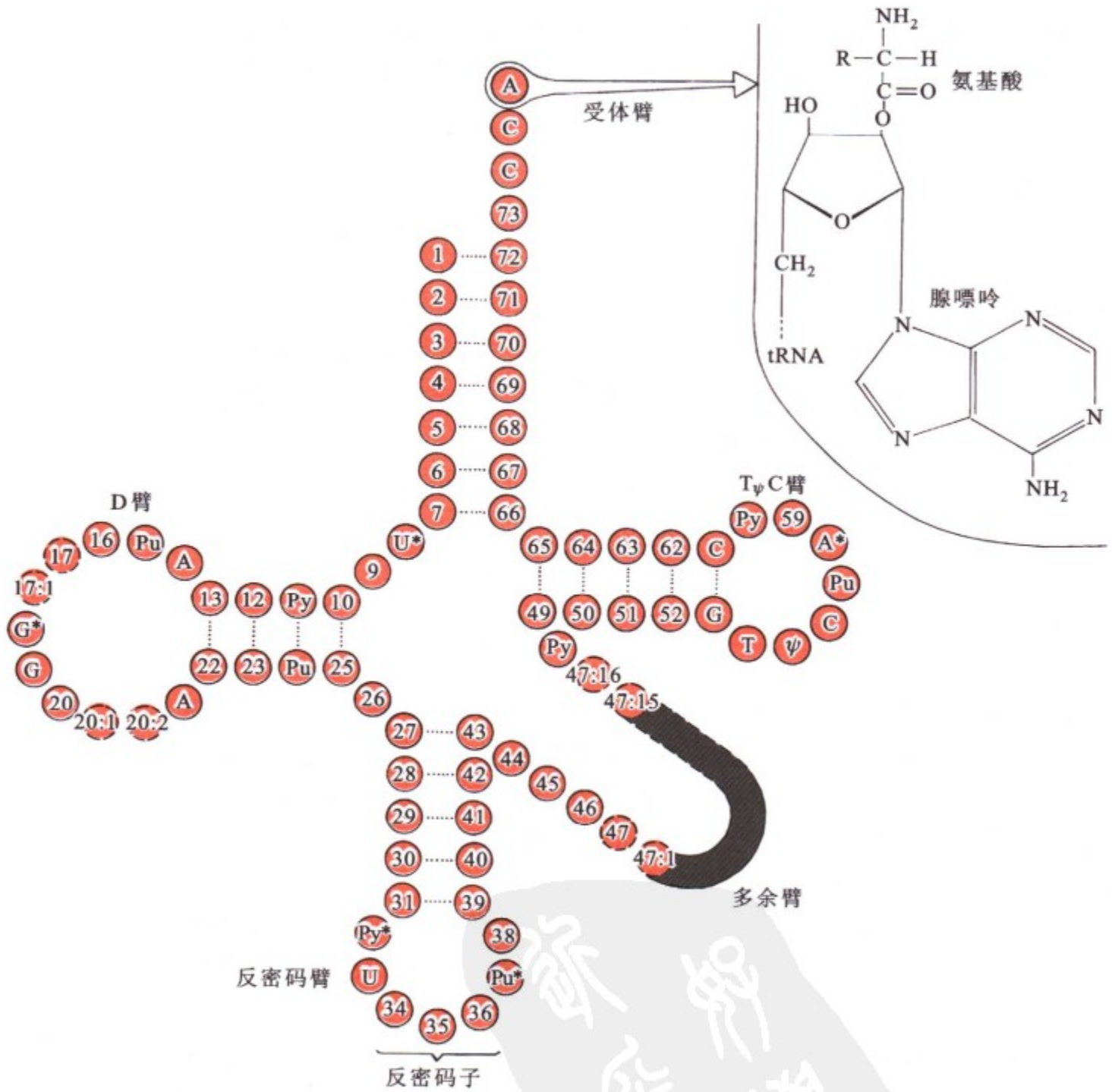



图 4-5 tRNA 的三叶草形二级结构

tRNA 分子中最大的变化发生在位于 $T_\psi C$ 和反密码子臂之间的多余臂(extra arm)上。根据多余臂的特性,又可以将 tRNA 分为两大类:第一类 tRNA 占所有 tRNA 的 75%,只含有一条仅为 3~5 个核苷酸的多余臂;第二类 tRNA 含有一条较大的多余臂,包括杆状结构上的 5 个核苷酸,套索结构上的 3~11 个核苷

酸。多余臂的生物学功能尚不清楚。

tRNA 的稀有碱基含量非常丰富，约有 70 余种。每个 tRNA 分子至少含有 2 个稀有碱基，最多有 19 个，多数分布在非配对区，特别是在反密码子 3′端邻近部位出现的频率最高，且大多为嘌呤核苷酸，这对于维持反密码子环的稳定性及密码子、反密码子之间的配对是很重要的。基因组研究发现，原核生物与真核生物细胞中所拥有的各种 *tRNA* 基因总数很不一样(表 4-6)。

表 4-6 几种不同生物中 *tRNA* 基因总数的比较

| 生物 | *tRNA* 基因总数 | 生物 | *tRNA* 基因总数 |
|---|---|---|---|
| 人类 | 497 | 酵母 | 273 |
| 线虫 | 584 | 大肠杆菌 | 86 |
| 果蝇 | 284 | | |

所有的 tRNA 都能够与核糖体的 A 位点(新进入的氨酰-tRNA 的结合位点)和P 位点(肽酰-tRNA 的结合位点)结合，此时，tRNA 分子三叶草型顶端突起部位通过密码子:反密码子的配对与 mRNA 相结合，而其 3′端恰好将所运转的氨基酸送到正在延伸的多肽上。此外，所有 tRNA(除了起始 tRNA 外)都能被翻译辅助因子 EF-Tu(原核生物)或 eEF1(真核生物)所识别而与核糖体相结合，起始 tRNA 能被原核生物的起始因子 IF-2 或真核生物的起始因子 eIF-2 所识别。

### 4.2.1 tRNA 的 L 形三级结构

有人在二级结构基础上做了酵母 $tRNA^{Phe}$、$tRNA^{fMet}$ 和大肠杆菌 $tRNA^{fMet}$、$tRNA^{Arg}$ 等的三级结构，发现都呈 L 形折叠(图 4-6)，而这种结构是靠氢键来维

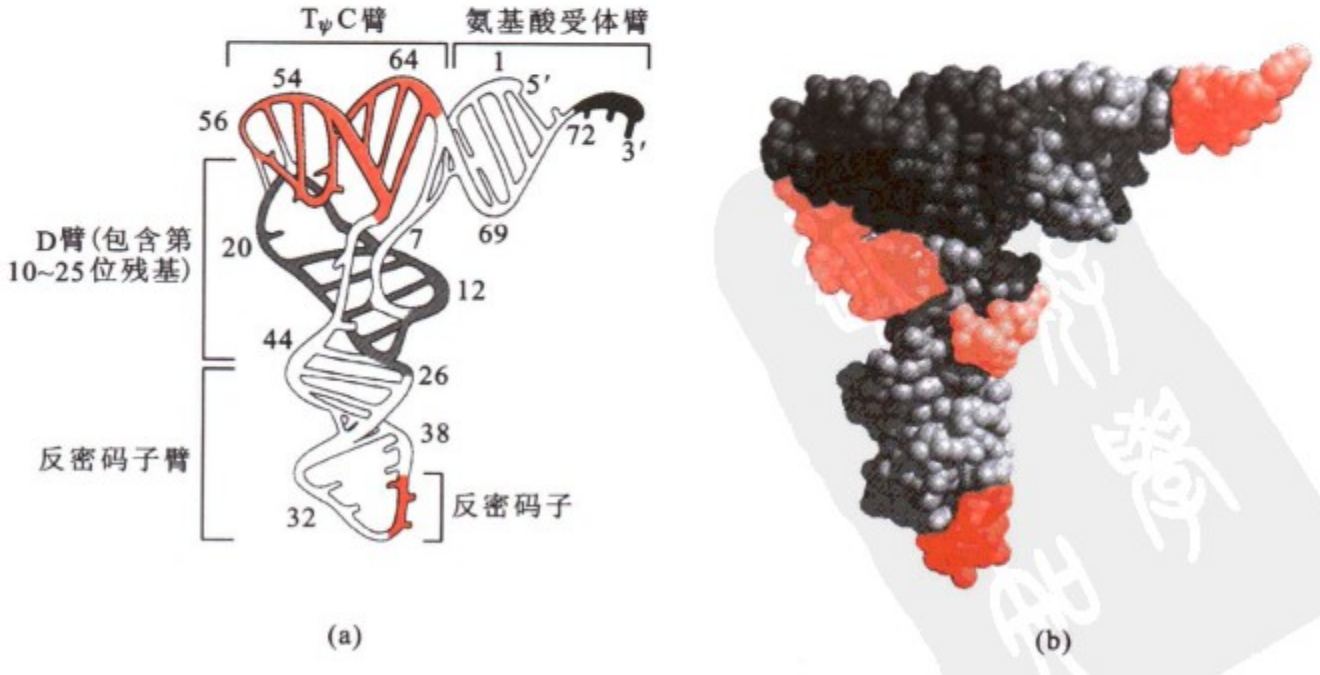



图 4-6 酵母 $tRNA^{Phe}$ 的三级结构示意图(根据 X 线衍射数据绘制) (a)和(b)表示用不同方法构建的模型。

持的。tRNA 的三级结构与 AA-tRNA 合酶对 tRNA 的识别有关。

tRNA 的 L 形三级结构，保留了二级结构中由于碱基互补而产生的双螺旋杆状结构，又通过分子重排创造了另一对双螺旋。受体臂和 $T_{\psi}C$ 臂的杆状区域构成了第一个双螺旋，D 臂和反密码子臂的杆状区域形成了第二个双螺旋，两个双螺旋上各有一个缺口。$T_{\psi}C$ 臂和 D 臂的套索状结构位于“L”的转折点。所以，受体臂顶端的碱基位于“L”的一个端点，反密码子臂的套索状结构生成了“L”的另一个端点(图 4-7)。

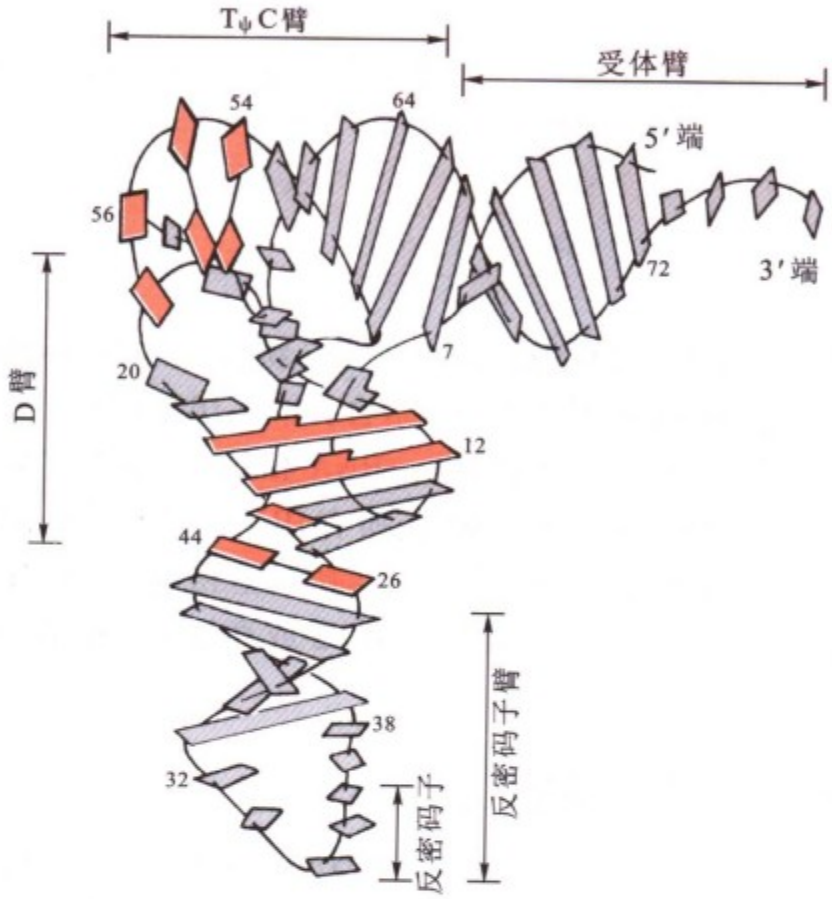



图 4-7 tRNA 的 L 形三级结构包含有两个双螺旋

tRNA 的三级结构主要由在二级结构中未配对碱基间形成氢键而引发的。在三叶草结构中的氢键被称为次级氢键，在三级结构中的就称为三级氢键。大部分恒定或半恒定核苷酸都参与三级氢键的形成。

tRNA 高级结构上的特点为我们提供了研究其生物学功能的重要线索，因为 tRNA 上所运载的氨基酸必须靠近位于核糖体大亚基上的多肽合成位点，而 tRNA 上的反密码子必须与小亚基上的 mRNA 相配对，所以分子中两个不同的功能基团是最大限度分离的。这个结构形式很可能满足了蛋白质合成过程中对 tRNA 的各种要求而成为 tRNA 的通式，研究证实 tRNA 的性质是由反密码子而不是它所携带的氨基酸所决定的。

### 4.2.2 tRNA 的功能

转录过程是信息从一种核酸分子(DNA)转移到另一种结构上极为相似的核酸分子（RNA）的过程，信息转移靠的是碱基配对。翻译阶段遗传信息从 mRNA 分子转移到结构极不相同的蛋白质分子，信息是以能被翻译成单个氨基

酸的三联密码子形式存在的,在这里起作用的是 tRNA 的解码机制。

根据 Crick 的接合体假说，氨基酸必须与一种接合体接合，才能被带到 RNA 模板的恰当位置上正确合成蛋白质。所以,氨基酸在合成蛋白质之前必须通过 AA-tRNA 合成酶活化,在消耗 ATP 的情况下结合到 tRNA 上,生成有蛋白质合成活性的 AA-tRNA。同时,AA-tRNA 的生成还涉及信息传递的问题,因为只有 tRNA 上的反密码子能与 mRNA 上的密码子相互识别并配对,而氨基酸本身不能识别密码子,只有结合到 tRNA 上生成 AA-tRNA,才能被带到 mRNA-核糖体复合物上,插入到正在合成的多肽链的适当位置上。

下面的实验证明模板 mRNA 只能识别特异的 tRNA 而不是氨基酸。$^{14}C$ 标记的半胱氨酸与 $tRNA^{Cys}$ 结合后生成 $^{14}C$-半胱氨酸-$tRNA^{Cys}$，经 Ni 催化可生成 $^{14}C$-Ala-$tRNA^{Cys}$,再把 $^{14}C$-Ala-$tRNA^{Cys}$ 加进含血红蛋白 mRNA、其他tRNA、氨基酸以及兔网织细胞核糖体的蛋白质合成系统中,结果发现 $^{14}C$-Ala-$tRNA^{Cys}$ 插入了血红蛋白分子通常由半胱氨酸占据的位置上,这表明在这里起识别作用的是 tRNA 而不是氨基酸。

### 4.2.3 tRNA 的种类

1. 起始 tRNA 和延伸 tRNA

有一类能特异地识别 mRNA 模板上起始密码子的 tRNA 叫起始 tRNA,其他 tRNA 统称为延伸 tRNA。起始 tRNA 具有独特的、有别于其他所有 tRNA 的结构特征。原核生物起始 tRNA 携带甲酰甲硫氨酸(fMet),真核生物起始 tRNA 携带甲硫氨酸 (Met)。原核生物中 Met-$tRNA^{fMet}$ 必须首先甲酰化生成 fMet-$tRNA^{fMet}$ 才能参与蛋白质的生物合成。

2. 同工 tRNA

由于一种氨基酸可能有多个密码子,因此有多个 tRNA 来识别这些密码子,即多个 tRNA 代表一种氨基酸，我们将几个代表相同氨基酸的 tRNA 称为同工 tRNA(cognate tRNA)。在一个同工 tRNA 组内,所有 tRNA 均专一于相同的氨酰-tRNA 合成酶。同工 tRNA 既要有不同的反密码子以识别该氨基酸的各种同义密码,又要有某种结构上的共同性,能被 AA-tRNA 合成酶识别。所以说,同工 tRNA 组内肯定具备了足以区分其他 tRNA 组的特异构造，保证合成酶能准确无误地加以选择。到目前为止,科学家还无法从一级结构上解释 tRNA 在蛋白质合成中的专一性。有证据说明,tRNA 的二级和三级结构对它的专一性起着举足轻重的作用。

3. 校正 tRNA

在蛋白质的结构基因中,一个核苷酸的改变可能使代表某个氨基酸的密码子变成终止密码子(UAG、UGA、UAA),使蛋白质合成提前终止,合成无功能的或无意义的多肽,这种突变称为无义突变(nonsense mutation),而无义突变的校正 tRNA 可通过改变反密码子区校正无义突变(表 4-7)。

错义突变(missense mutation)是由于结构基因中某个核苷酸的变化使一种氨基酸的密码变成另一种氨基酸的密码。错义突变的校正 tRNA 通过反密码子

表 4-7 大肠杆菌无义突变的校正 tRNA

| 位点 | tRNA | 野生型 | | 校正基因 | |
|---|---|---|---|---|---|
| | | 识别密码 | 反密码 | 识别密码 | 反密码 |
| | | | ← | | ← |
| supD | Ser | UCG | CGA | UAG | CUA |
| | | | ← | | ← |
| supE | Gln | CAG | CUG | UAG | CUA |
| | | | ← | | ← |
| supC | Tyr | UAC | GUA | UAG | CUA |
| | | | ← | | ← |
| supG | Lys | AAA | UUU | UAA | UUA |
| | | | ← | | ← |
| supU | Trp | UGG | CCA | UGA | UCA |

区的改变把正确的氨基酸加到肽链上,合成正常的蛋白质。如某大肠杆菌细胞色氨酸合成酶中的一个甘氨酸密码子 GGA 错义突变成 AGA(编码精氨酸),指导合成错误的多肽链。甘氨酸校正 tRNA 的校正基因突变使其反密码子从 CCU 变成 UCU,它仍然是甘氨酸的反密码子但不结合 GGA 而能与突变后的 AGA 密码子结合,把正确的氨基酸(甘氨酸)放到 AGA 所对应的位置上。

校正 tRNA 在进行校正过程中必须与正常的 tRNA 竞争结合密码子,无义突变的校正 tRNA 必须与释放因子竞争识别密码子,错义突变的校正 tRNA 必须与该密码的正常 tRNA 竞争,这些都会影响校正的效率。所以,某个校正基因的效率不仅决定于反密码子与密码子的亲和力,也决定于它在细胞中的浓度及竞争中的其他参数。一般说来,校正效率不会超过 50%。无义突变的校正基因 tRNA 不仅能校正无义突变,也会抑制该基因 3′端正常的终止密码子,导致翻译过程的通读,合成更长的蛋白质,这种蛋白质过多就会对细胞造成伤害。同样,一个基因错义突变的校正也可能使另一个基因错误翻译,因为如果一个校正基因在突变位点通过取代一种氨基酸的方式校正了一个突变,它也可以在另一位点这样做,从而在正常位点上引入新的氨基酸。

### 4.2.4 氨酰-tRNA 合成酶

氨酰-tRNA 合成酶(AA-tRNA synthetase)是一类催化氨基酸与 tRNA 结合的特异性酶,其反应式如下:

$$AA+tRNA+ATP \rightarrow AA\text{-}tRNA+AMP+PP_i$$

它实际上包括两步反应:第一步是氨基酸活化生成酶-氨酰腺苷酸复合物。

$$AA+ATP+酶(E) \rightarrow E\text{-}AA\text{-}AMP+PP_i$$

第二步是氨酰基转移到 tRNA 3′端腺苷残基的 2′或 3′-羟基上。

$$E\text{-}AA\text{-}AMP+tRNA \rightarrow AA\text{-}tRNA+E+AMP$$

蛋白质合成的真实性主要决定于 tRNA 能否把正确的氨基酸放到新生多肽

链的正确位置上，而这一步主要决定于 AA-tRNA 合成酶是否能使氨基酸与对应的 tRNA 相结合。AA-tRNA 合成酶既要能识别 tRNA，又要能识别氨基酸，它对两者都具有高度的专一性。不同的 tRNA 有不同的碱基组成和空间结构，易被特异性氨酰-tRNA 合成酶所识别，困难的是这些酶怎么去识别结构上非常相似的氨基酸。例如，异亮氨酸只比缬氨酸多一个甲烯基团，而异亮氨酸-tRNA 合成酶对异亮氨酸的亲和力比对缬氨酸大 225 倍。考虑到体内缬氨酸的浓度比异亮氨酸高 5 倍，缬氨酸被错误活化渗入到异亮氨酸位点上去的概率应该是 1/40，但实际上错误频率(误差率)只有约 $1.5\times10^{-5}$，表明还存在其他重要的校正手段以增加蛋白质合成的真实性。有科学家发现，被错误活化的缬氨酸不会被结合到 $tRNA^{Ile}$ 上，而是被酶本身水解，即活化阶段产生的误差在后一阶段被再次校正了。

## 4.3 核糖体

生物细胞内，核糖体像一个能沿 mRNA 模板移动的工厂，执行着蛋白质合成的功能。它是由几十种蛋白质和几种核糖体 RNA(ribosomal RNA，rRNA)组成的亚细胞颗粒。一个细菌细胞内约有 20 000 个核糖体，而真核细胞内可达 $10^6$ 个，在未成熟的蟾蜍卵细胞内则高达 $10^{12}$ 个。这些颗粒既可以游离状态存在于细胞内，也可与内质网结合，形成微粒体。核糖体和它的辅助因子为蛋白质生物合成提供了必要的条件。

运载有肽链起始或延伸必需氨基酸的 AA-tRNA，往往以令人难以置信的速度进入核糖体，在起始或延伸因子的作用下，与 mRNA 模板和延伸中的肽链相互作用，卸去所载氨基酸后立即退出核糖体，以保证新一轮合成反应的顺利进行。

核糖体蛋白约占原核细胞总蛋白量的 10%，占细胞内总 RNA 量的 80%(表 4-8)。在真核细胞内，核糖体 RNA 所占的比例虽然有所下降，但仍然占总 RNA 的绝大部分，是细胞总蛋白的一个重要组成部分。无论原核或真核细胞内核糖体的含量都是与细胞蛋白质合成活性直接相关的。

表 4-8 核糖体及其他组分在大肠杆菌细胞内的分布

| 组分 | 占细胞总量 | 细胞内数量 |
|---|---|---|
| 细胞壁 | 10% | 1 |
| 细胞膜 | 10% | 2 |
| DNA | 2% | 1 |
| mRNA | 2% | $3.5\times10^3$ |
| tRNA | 3% | $1.6\times10^5$ |
| rRNA | 21% | $8\times10^5$ |
| 核糖体蛋白 | 9% | $2\times10^4$ |
| 可溶性蛋白 | 40% | $10^6$ |
| 小分子 | 3% | $7.5\times10^6$ |

在真核生物中,大多数正在进行蛋白质合成的核糖体都不是在细胞质内自由漂浮,而是直接或间接地与细胞骨架结构有关联或者与内质网膜结构相连的(图 4-8)。细菌核糖体大都通过与 mRNA 的相互作用,被固定在核基因组上。

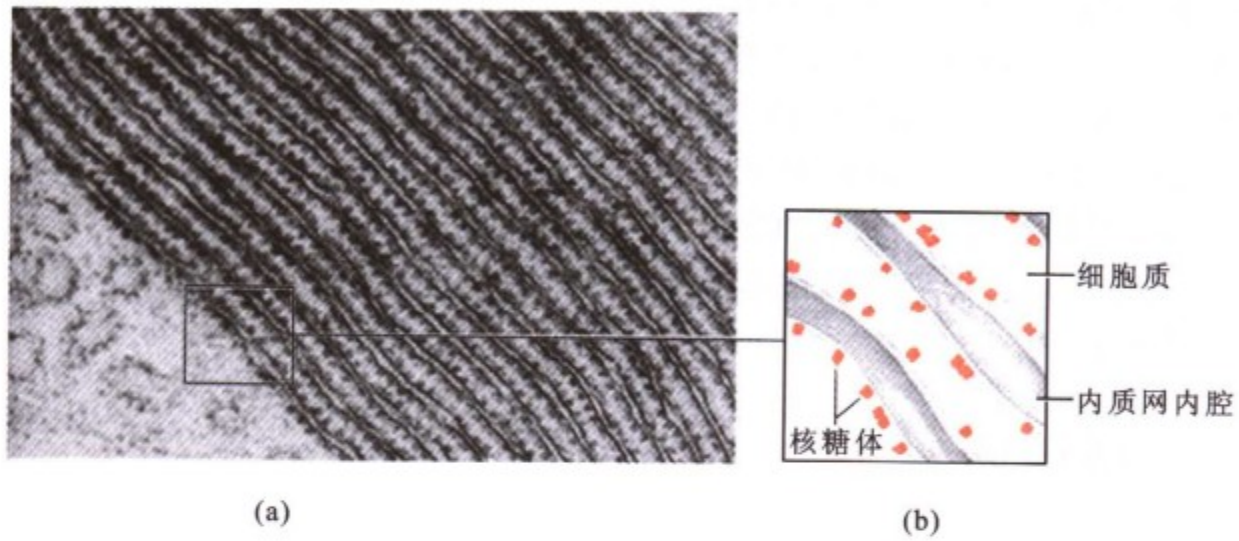



**图 4-8 结合在内质网上的核糖体** (a) 电镜下看到的胰腺细胞糙面内质网;(b) 局部放大后的图像。

核糖体包括两个亚基,大亚基约为小亚基相对分子质量的两倍。每个亚基包含一个主要的 rRNA 成分和许多不同功能的蛋白质分子,这些分子大都以单拷贝的形式存在。大亚基除了含有主要 rRNA 组分外,还有一些相对分子质量较小的 RNA。核糖体亚基中的主要 *rRNA* 基因虽然有许多个拷贝,但其序列十分保守,暗示 rRNA 序列在组成功能核糖体的过程中有着重要的作用。表 4-9 是大肠杆菌核糖体的基因组成成分,从已知的大肠杆菌 rRNA 和蛋白质序列来看,细菌的核糖体是完全相同的。

**表 4-9 大肠杆菌核糖体基本成分**

| | 核糖体 | 小亚基 | 大亚基 |
|---|---|---|---|
| 沉降系数 | 70S | 30S | 50S |
| 总体相对分子质量 | $2.52 \times 10^6$ | $9.30 \times 10^5$ | $1.59 \times 10^6$ |
| 主要 rRNA(碱基数) | | 16S(1 541) | 23S(2 004) |
| 主要 rRNA(碱基数) | | | 5S(120) |
| RNA 相对分子质量 | $1.66 \times 10^6$ | $5.60 \times 10^5$ | $1.10 \times 10^6$ |
| RNA 所占比例 | 66% | 60% | 70% |
| 蛋白质数量 | | 21 | 36 |
| 蛋白质相对分子质量 | $8.57 \times 10^5$ | $3.70 \times 10^5$ | $4.87 \times 10^5$ |
| 蛋白质所占比例 | 34% | 40% | 30% |

### 4.3.1 核糖体的结构

1. 核糖体由大小两个亚基组成

核糖体是一个致密的核糖核蛋白颗粒,可以解离为两个亚基,每个亚基都含有一个相对分子质量较大的 rRNA 和许多不同的蛋白质分子。这些大分子 rRNA 能在特定位点与蛋白质结合,从而完成核糖体不同亚基的组装。原核生物核糖体由约 2/3 的 RNA 及 1/3 的蛋白质组成。真核生物核糖体中 RNA 占 3/5,蛋白质占 2/5。原核生物、真核生物细胞质及细胞器中的核糖体存在着很大差异(表 4-10)。

表 4-10 几种不同生物核糖体及 rRNA 的组成

| 核糖体 | 来 源 | 大亚基 | | 小亚基 | |
|---|---|---|---|---|---|
| | | 沉降系数 | RNA | 沉降系数 | RNA |
| 80 S | 脊椎动物 | 60 S | 20~29 S | 40 S | 18 S |
| | | | 5 S | | |
| | | | 5.8 S | | |
| 80 S | 无脊椎动物、植物 | 60 S | 25 S | 40 S | 16~18 S |
| | | | 5 S | | |
| | | | 5.8 S | | |
| 70 S | 原核生物 | 50 S | 23 S | 30 S | 16 S |
| 55 S | 脊椎动物线粒体 | 40 S | 16~17 S | 30 S | 10~13 S |

2. 核糖体蛋白(ribosomal protein,r-蛋白)

核糖体上有多个活性中心,每个中心都由一组特殊的核糖体蛋白质(r-protein)构成。虽然有些蛋白质本身具有催化功能,但若将它们从核糖体上分离出来,催化功能就会完全消失。所以,核糖体是一个许多酶的集合体,单个酶或蛋白只有在这个总体结构内才拥有催化活性,它们在这一结构中共同承担了蛋白质生物合成的任务。核糖体中许多蛋白质(可能还包括 rRNA)的主要功能可能就是建立这种总体结构,使各个活性中心处于适当的相互协调的关系之中。

已知大肠杆菌核糖体小亚基由 21 种 r-蛋白组成,分别用 $S^1$……$S^{21}$ 表示,大亚基由 36 种 r-蛋白组成,分别用 $L^1$……$L^{36}$ 表示。真核生物细胞核糖体蛋白质中,大亚基含有 49 种 r-蛋白,小亚基有 33 种 r-蛋白,它们的相对分子质量在 $8\times10^3$~$4.0\times10^4$。

对核糖体蛋白质的二级和三级结构也有不少研究。图 4-9 是在高倍电镜下得到的原核生物 70S 核糖体大、小亚基相结合的模型,核糖体分子可容纳两个 tRNA 和约 40 bp 长的 mRNA。

3. 核糖体 RNA

核糖体内的 rRNA 不仅是核糖体的重要结构成分,也是核糖体发挥生理功能的重要元件。下面详细介绍各类 rRNA 的组成和功能特点。

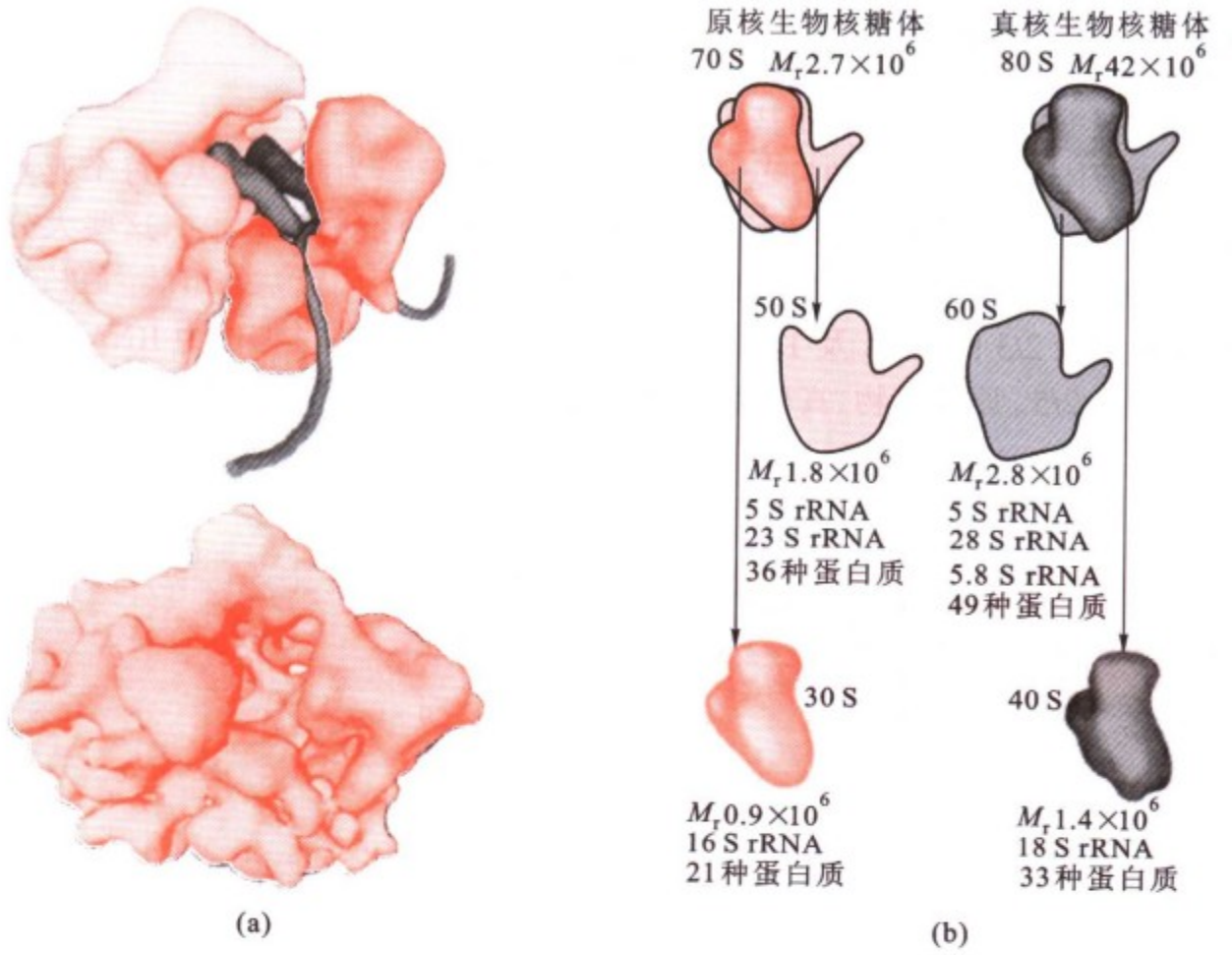



**图 4-9 核糖体结构模型及原核与真核细胞核糖大小亚基比较** (a) 电镜数据模式图；大亚基位于整个分子的左侧，细绳代表 mRNA，位于两个亚基之间的是两个 tRNA 分子；(b) 原核生物 70 S 核糖体(相对分子质量2.7×10⁶)和真核生物 80 S 核糖体(相对分子质量 4.2×10⁶)。

(1) 5S rRNA 细菌 5S rRNA 含有 120 个核苷酸(革兰阴性菌)或 116 个核苷酸(革兰阳性菌)。5S rRNA 有两个高度保守的区域,其中一个区域含有保守序列 CGAAC,这是与 tRNA 分子 $T_\psi C$ 环上的 $GT_\psi CG$ 序列相互作用的部位,是 5S rRNA 与 tRNA 相互识别的序列。另一个区域含有保守序列 GCGCCGAAUGGUAGU,与 23S rRNA 的中一段序列互补,这是 5S rRNA 与 50S 核糖体大亚基相互作用的位点,在结构上有其重要性。

(2) 16S rRNA 16S rRNA 在蛋白质的合成中起着积极作用,它与 mRNA、50S 亚基和 P 位和 A 位的 tRNA 的反密码子直接作用。其长度在 1 475~1 544 个核苷酸,含有少量修饰碱基。该分子虽然可被分成几个区,但全部压缩在 30 S 小亚基内。16S rRNA 的结构十分保守，其中 3′端一段 ACCUCCUUA 的保守序列,与 mRNA 5′端翻译起始区富含嘌呤的序列互补。在 16S rRNA 靠近 3′端处还有一段与 23S rRNA 互补的序列,在 30S 与 50S 亚基的结合中起作用。

(3) 23S rRNA 23S *rRNA* 基因的一级结构包括 2 904 个核苷酸，在大肠杆菌 23S rRNA 第 1 984~2 001 核苷酸，存在一段能与 $tRNA^{Met}$ 序列互补的片段,表明核糖体大亚基 23S rRNA 可能与 $tRNA^{Met}$ 的结合有关。在 23S rRNA 靠近 5′端(143~157 位核苷酸)有一段 12 个核苷酸的序列与 5S rRNA 上第 72~83

位核苷酸互补,表明在 50S 大亚基上这两种 RNA 之间可能存在相互作用。核糖体 50S 大亚基上约有 20 种蛋白质能不同程度地与 23S rRNA 结合。

(4) 5.8S rRNA　5.8S rRNA 是真核生物核糖体大亚基特有的 rRNA,长度为 160 个核苷酸,含有修饰碱基。它还含有与原核生物 5S rRNA 中的保守序列CGAAC 相同的序列,可能是与 tRNA 作用的识别序列,这说明 5.8S rRNA 可能与原核生物的 5S rRNA 具有相似的功能。表 4-11 是人类基因组中主要 *rRNA* 基因拷贝数分析,其中 18S 和 5.8S *rRNA* 基因的数量大大低于原先的估计。

表 4-11　人类基因组中主要 *rRNA* 基因拷贝数分析

| *rRNA* 基因 | 预计拷贝数 | 实际拷贝数 | 相关基因拷贝数 |
|---|---|---|---|
| 18S | 150 ~ 200 | 0 | 40 |
| 5.8S | 150 ~ 200 | 1 | 11 |
| 28S | 150 ~ 200 | 0 | 181 |
| 5S | 200 ~ 300 | 4 | 520 |

(5) 18S rRNA　酵母 18S rRNA 由 1 789 个核苷酸组成,它的 3′端与大肠杆菌 16S rRNA 有广泛的同源性。其中酵母 18S rRNA、大肠杆菌 16S rRNA 和人线粒体 12S rRNA 在 3′端有 50 个核苷酸序列相同。

28S rRNA 长度约在 3 890~4 500 bp,目前还不清楚该 rRNA 的功能。从上述讨论中可以看出 rRNA 与 tRNA 及 mRNA 之间的相互关系以及不同的 rRNA 之间的关系,这种关系是建立在序列互补或同源的基础之上的。

4. 核糖体有 3 个 tRNA 结合位点

氨基酸由氨酰-tRNA 运送到核糖体上并通过这个 tRNA 与携带上一个氨基酸的 tRNA 的相互作用,将新的氨基酸加到正在生长的新生蛋白质链上。核糖体上有 3 个 tRNA 结合位点,分别称为 A、P 和 E 位点(图 4-10)。其中,A 位点是新到来的氨酰-tRNA 的结合位点;P 位点是肽酰-tRNA 的结合位点;E 位点是延伸过程中的多肽链转移到氨酰-tRNA 上释放 tRNA 的位点,即去氨酰-tRNA 通过 E 位点脱出,被释放到核糖体外的细胞质中去。所以 tRNA 的移动顺序是从 A 位到 P 位再到 E 位,通过密码子与反密码子之间的相互作用保证反应正向进行而不会倒转。

图 4-10　核糖体的 3 个 tRNA 结合位点,即 A、P 和 E 位点

由于 tRNA 的氨酰末端被定位到大亚基上,而另一端的反密码子与结合小亚基的 mRNA 相互识别,所以,每一个 tRNA 结合位点都横跨核糖体的两个亚基,位于大、小亚基的交界面。

### 4.3.2 核糖体的功能

核糖体存在于每个进行蛋白质合成的细胞中。虽然在不同生物体内其大小有别,但组织结构基本相同,而且执行的功能也完全相同(图 4-11)。在多肽合成过程中,不同的 tRNA 将相应的氨基酸带到蛋白质合成部位,并与 mRNA 进行专一性的相互作用,以选择对信息专一的 AA-tRNA。核糖体还必须能同时容纳另一种携带肽链的 tRNA,即肽酰-tRNA(peptidyl-tRNA),并使之处于肽键易于生成的位置上。

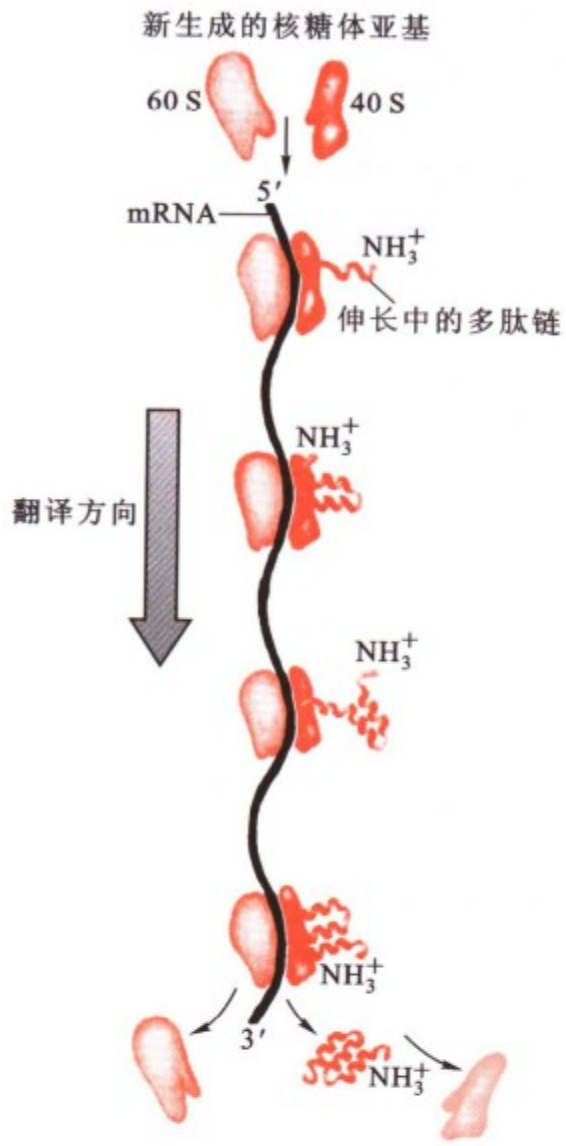



图 4-11 真核细胞中发现的多聚核糖体现象

核糖体包括多个活性中心,即 mRNA 结合部位、结合或接受 AA-tRNA 部位(A 位)、结合或接受肽酰-tRNA 的部位(P 位)、肽基转移部位及形成肽键的部位(转肽酶中心)。此外,还应有负责肽链延伸的各种延伸因子的结合位点。

核糖体小亚基负责对模板 mRNA 进行序列特异性识别,如起始部分的识别、密码子与反密码子的相互作用等,mRNA 的结合位点也在此亚基上。大肠杆菌中与翻译的真实性有关的蛋白质 S4 及 S12 也属此亚基。大亚基负责携带氨基酸及 tRNA 的功能,包括肽键的形成、AA-tRNA、肽酰-tRNA 的结合等。

核糖体在体内及体外都可解离为亚基或结合成 70S/80S 的颗粒。在翻译的起始阶段需要游离的亚基,随后才结合成 70S/80S 颗粒,开始翻译进程。大肠杆菌中的起始因子 IF-3 与小亚基结合,阻碍了后者与大亚基的结合,从而控制翻译起始过程。体外反应体系中,核糖体的解离或结合取决于离子浓度。在大肠杆菌内,$Mg^{2+}$浓度在 $10^{-3}$mol/L 以下时,70S 解离为亚基,浓度达 $10^{-2}$mol/L 时则形成稳定的 70S 颗粒。mRNA 链上一般每 40 个核苷酸有一个核糖体。肽链释放后,核糖体脱离 mRNA 解聚成亚基,直接参与另一轮蛋白质的合成或两个亚基结合生成稳定的核糖体,不参与蛋白质合成。

## 4.4 蛋白质合成的生物学机制

蛋白质(protein)一词从希腊文“proteios”衍生而来的,是指按重要性排列居于第一位的东西。尽管现已证明核酸是生命体内最基本的物质,因为蛋白质的合成和结构最终都取决于核酸,但蛋白质仍是生物活性物质中最重要的大分子组分,生物有机体的遗传学特性仍然要通过蛋白质来得到表达。

蛋白质的生物合成包括氨基酸活化,肽链的起始、伸长、终止以及新合成多肽链的折叠和加工,现将各阶段的必需成分列于表 4-12。

表 4-12 蛋白质合成各阶段的主要成分简表

| 阶 段 | 必需组分 |
| --- | --- |
| 氨基酸的活化 | 20 种氨基酸<br>20 种氨酰 - tRNA 合成酶<br>20 种或更多的 tRNA<br>ATP, $Mg^{2+}$ |
| 肽链的起始 | mRNA<br>*N* - 甲酰甲硫氨酰 - tRNA<br>mRNA 上的起始密码子(AUG)<br>核糖体小亚基<br>核糖体大亚基<br>GTP, $Mg^{2+}$<br>起始因子(IF - 1, IF - 2, IF - 3) |
| 肽链的延伸 | 功能核糖体(起始复合物)<br>AA - tRNA<br>伸长因子<br>GTP, $Mg^{2+}$<br>肽基转移酶 |
| 肽链的终止 | GTP<br>mRNA 上的终止密码子<br>释放因子(RF - 1, RF - 2, RF - 3) |
| 折叠和加工 | 参与起始氨基酸的切除、修饰等加工过程的酶 |

## 4.4.1 氨基酸的活化

蛋白质的生物合成是以氨基酸作为基本建筑材料的，且只有与 tRNA 相结合的氨基酸才能被准确地运送到核糖体中,参与多肽链的起始或延伸。表 4-13 是蛋白质中全部 20 种氨基酸的主要特征分析,而图 4-12 是这些氨基酸的结构式。氨基酸必须在氨酰-tRNA 合成酶的作用下生成活化氨基酸——AA-tRNA。研究发现,至少存在 20 种以上具有氨基酸专一性的氨酰-tRNA 合成酶,能够识别并通过氨基酸的羧基与 tRNA 3′端腺苷酸核糖基上 3′—OH 缩水形成二酯键。同一氨酰-tRNA 合成酶具有把相同氨基酸加到两个或更多个带有不同反密码子 tRNA 分子上的功能。

tRNA 与相应氨基酸的结合是蛋白质合成中的关键步骤，因为只要 tRNA 携带了正确的氨基酸,多肽合成的准确性就相对有了保障。

在细菌中,起始氨基酸是甲酰甲硫氨酸,所以,与核糖体小亚基相结合的是 *N*-甲酰甲硫氨酰-$tRNA^{fMet}$,它由以下两步反应合成:

表 4-13 参与蛋白质合成的 20 种氨基酸主要特征分析

| 氨基酸 | 简称 | | 相对分子质量 | pI | 疏水性指数* | 出现频率 |
|---|---|---|---|---|---|---|
| 非极性 | | | | | | |
| 带脂肪族 R 基团 | | | | | | |
| 甘氨酸 | Gly | G | 75 | 5.97 | -0.4 | 7.2 |
| 丙氨酸 | Ala | A | 89 | 6.01 | 1.8 | 7.8 |
| 缬氨酸 | Val | V | 117 | 5.97 | 4.2 | 6.6 |
| 亮氨酸 | Leu | L | 131 | 5.98 | 3.8 | 9.1 |
| 异亮氨酸 | Ile | I | 131 | 6.02 | 4.5 | 5.3 |
| 甲硫氨酸 | Met | M | 149 | 5.74 | 1.9 | 2.3 |
| 带芳香族 R 基团 | | | | | | |
| 苯丙氨酸 | Phe | F | 165 | 5.48 | 2.8 | 3.9 |
| 酪氨酸 | Tyr | Y | 181 | 5.66 | -1.3 | 3.2 |
| 色氨酸 | Trp | W | 204 | 5.89 | -0.9 | 1.4 |
| 极性 | | | | | | |
| R 基团呈中性 | | | | | | |
| 丝氨酸 | Ser | S | 105 | 5.68 | -0.8 | 6.8 |
| 脯氨酸 | Pro | P | 115 | 6.48 | 1.6 | 5.2 |
| 苏氨酸 | Thr | T | 119 | 5.87 | -0.7 | 5.9 |
| 半胱氨酸 | Cys | C | 121 | 5.07 | 2.5 | 1.9 |
| 天冬酰胺 | Asn | N | 132 | 5.41 | -3.5 | 4.3 |
| 谷氨酰胺 | Gln | Q | 146 | 5.65 | -3.5 | 4.2 |
| R 基团带正电荷 | | | | | | |
| 赖氨酸 | Lys | K | 146 | 9.74 | -3.9 | 5.9 |
| 组氨酸 | His | H | 155 | 7.59 | -3.2 | 2.3 |
| 精氨酸 | Arg | R | 174 | 10.76 | -4.5 | 5.1 |
| R 基团带负电荷 | | | | | | |
| 天冬氨酸 | Asp | D | 133 | 2.77 | -3.5 | 5.3 |
| 谷氨酸 | Glu | E | 147 | 3.22 | -3.5 | 6.3 |

* 负值表示亲水性,正值表示疏水性。

$$Met+tRNA^{fMet}+ATP \rightarrow Met\text{-}tRNA^{fMet}+AMP+PP_i$$

然后,由甲酰基转移酶转移一个甲酰基到 Met 的氨基上。

$$N^{10}\text{-甲酰四氢叶酸}+Met\text{-}tRNA^{fMet} \rightarrow \text{四氢叶酸}+fMet\text{-}tRNA^{fMet}$$

真核生物中,任何一个多肽合成都是从生成甲硫氨酰-$tRNAi^{Met}$ 开始的,因为甲硫氨酸的特殊性,所以体内存在两种 $tRNA^{Met}$。只有甲硫氨酰-$tRNAi^{Met}$ 才能与 40S 小亚基相结合,起始肽链合成,普通 $tRNA^{Met}$ 中携带的甲硫氨酸只能被掺入正在延伸的肽链中。

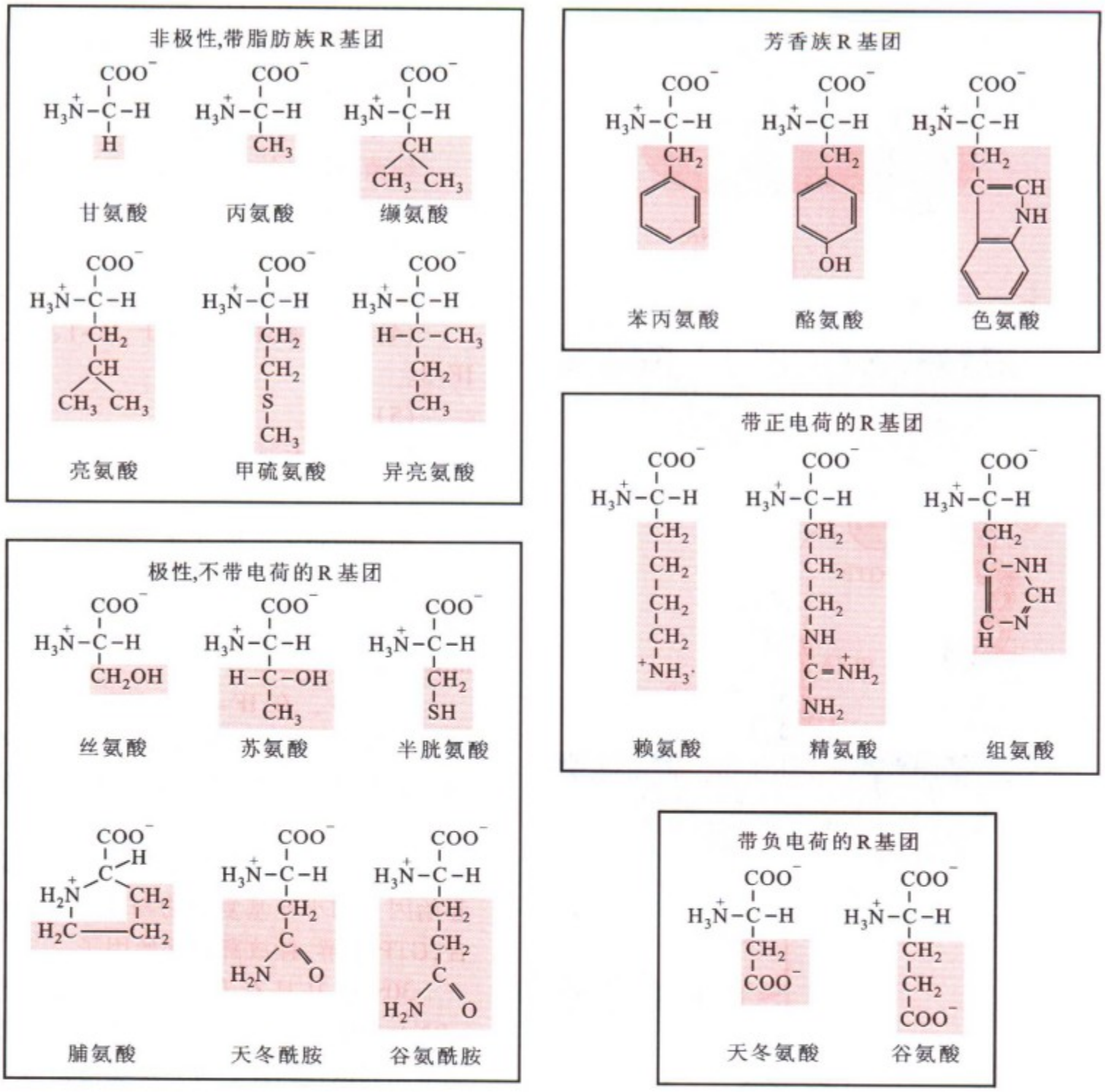



图 4-12 20 种氨基酸的结构

## 4.4.2 翻译的起始

蛋白质合成的起始需要核糖体大、小亚基,起始 tRNA 和几十个蛋白因子的参与,在模板 mRNA 编码区 5′端形成核糖体–mRNA–起始 tRNA 复合物并将甲酰甲硫氨酸放入核糖体 P 位点。

原核生物的起始 tRNA 是 $fMet\text{-}tRNA^{fMet}$,真核生物是 $Met\text{-}tRNA^{Met}$。原核生物中 30S 小亚基首先与 mRNA 模板相结合,再与 $fMet\text{-}tRNA^{fMet}$ 结合,最后与 50S 大亚基结合。而在真核生物中,40S 小亚基首先与 $Met\text{-}tRNA^{Met}$ 相结合,再与模板 mRNA 结合,最后与 60S 大亚基结合生成 80S·mRNA·$Met\text{-}tRNA^{Met}$ 起始复合物。起始复合物的生成除了需要 GTP 提供能量外,还需要 $Mg^{2+}$、$NH_4^+$ 及 3 个起始因

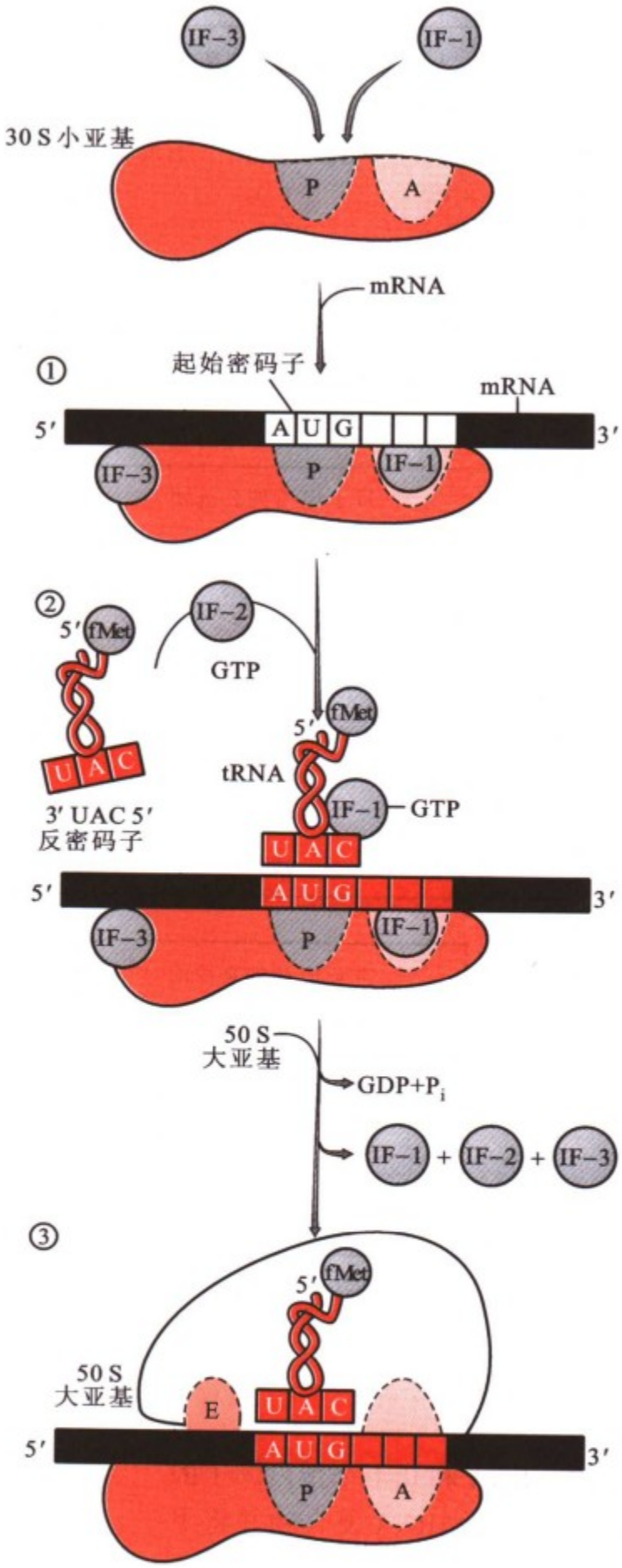



图 4-13 翻译起始复合物的形成

子(IF-1、IF-2、IF-3)。起始因子与 30S 小亚基的结合较为松散,用 1 mol/L $NH_4Cl$ 处理即可使之游离。

1. 原核生物翻译的起始

细菌中翻译的起始需要如下 7 种成分:

(1) 30S 小亚基。

(2) 模板 mRNA。

(3) fMet-tRNA$^{fMet}$。

(4) 三个翻译起始因子,IF-1、IF-2 和 IF-3。

(5) GTP。

(6) 50S 大亚基。

(7) $Mg^{2+}$。

翻译起始又可被分成三步(图 4-13)。

第一步,30S 小亚基首先与翻译起始因子 IF-1,IF-3 结合,通过 SD 序列与 mRNA 模板相结合。

第二步,在 IF-2 和 GTP 的帮助下,fMet-tRNA$^{fMet}$ 进入小亚基的 P 位,tRNA 上的反密码子与 mRNA 上的起始密码子配对。

第三步,带有 tRNA,mRNA,三个翻译起始因子的小亚基复合物与 50S 大亚基结合,GTP 水解,释放翻译起始因子。

30S 亚基具有专一性的识别和选择 mRNA 起始位点的性质,而 IF-3 能协助该亚基完成这种选择。研究发现,30S 亚基通过其 16S rRNA 的 3′端与 mRNA5′端起始密码子上游碱基配对结合。Shine 及 Dalgarno 等证明几乎所有原核生物mRNA 上都有一个 5′-AGGAGGU-3′序列,这个富嘌呤区被命名为 SD 序列(Shine-Dalgarno sequence),它与 30S 亚基上 16S rRNA 3′端的富嘧啶区序列 5′-GAUCACCUCCUUA-3′相互补(图 4-14)。

各种 mRNA 的核糖体结合位点中能与 16S rRNA 配对的核苷酸数目及这些核苷酸到起始密码子之间的距离是不一样的,反映了起始信号的不均一性。一般说来,相互补的核苷酸越多,30S 亚基与 mRNA 起始位点结合的效率也越高。互补的核苷酸与 AUG 之间的距离也会影

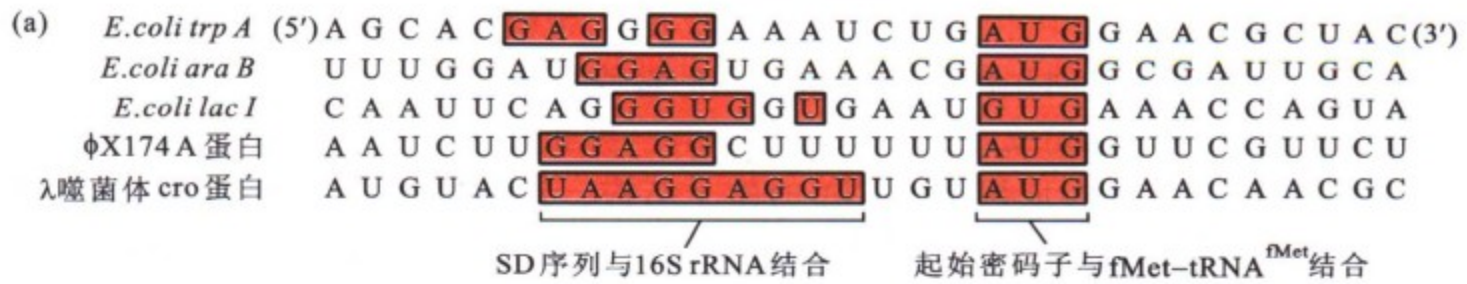


图 4-14 细菌 mRNA 分子上往往存在一个与 16S rRNA 3′端相互补的 SD序列

响 mRNA-核糖体复合物的形成及其稳定性。

细菌核糖体上一般存在三个与氨酰-tRNA 结合的位点,即 A 位点(aminoacyl site),P 位点(peptidyl site)和 E 位点(exit site)。只有 fMet-tRNA$^{fMet}$ 能与第一个 P 位点相结合,其他所有 tRNA 都必须通过 A 位点到达 P 位点,再由 E 位点离开核糖体。

已知 IF-2 对于 30S 起始复合物与 50S 亚基的连接是必需的,而 IF-1 则在 70S 起始复合物生成后促进 IF-2 的释放,从而完成蛋白质合成的起始过程。

2. 真核生物翻译的起始

真核生物蛋白质生物合成的起始机制与原核生物基本相同,其差异主要是核糖体较大,有较多的起始因子参与,其 mRNA 具有 m$^7$GpppNp 帽子结构,Met-tRNA$^{Met}$ 不甲酰化,mRNA 分子 5′端的“帽子”和 3′端的多聚 A 都参与形成翻译起始复合物。

有实验说明帽子结构能促进起始反应,因为核糖体上有专一位点或因子识别 mRNA 的帽子,使 mRNA 与核糖体结合(图 4-15)。帽子在 mRNA 与 40S 亚

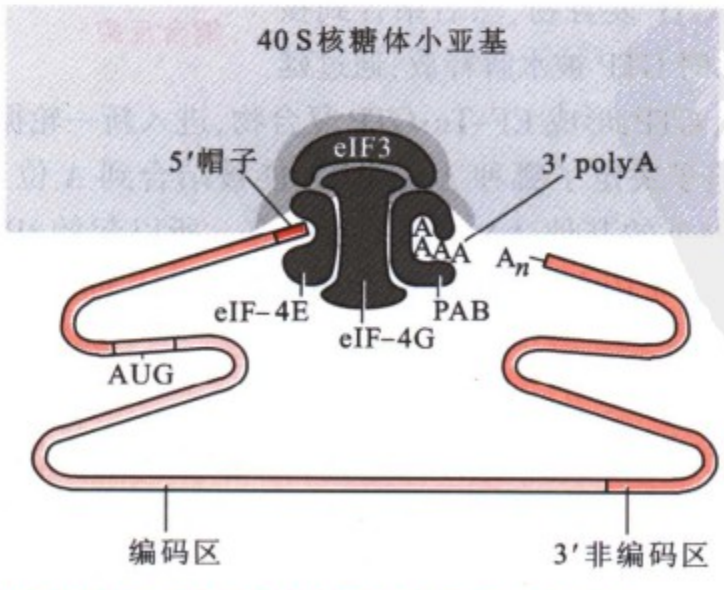



图 4-15 真核生物翻译起始复合物的形成

基结合过程中还起稳定作用。实验表明,带帽子的 mRNA 5′端与 18S rRNA 的 3′端序列之间存在不同于 SD 序列的碱基配对型相互作用。

40S 起始复合物形成过程中有一种蛋白因子——帽子结合蛋白 (eIF-4E),能专一地识别 mRNA 的帽子结构,与 mRNA 的 5′端结合生成蛋白质-mRNA 复合物,并利用该复合物对 eIF-3 的亲和力与含有 eIF-3 的 40S 亚基结合。

除了帽子结构以外,40S 小亚基还能识别 mRNA 上的起始密码子 AUG。Kozak 等提出了一个“扫描模型”来解释 40S 亚基对 mRNA 起始密码子的识别作用。按照这个模型,40S 小亚基先结合在 mRNA 5′端的任何序列上,然后沿 mRNA 移动直至遇到 AUG 发生较为稳定的相互作用, 最后与 60S 亚基一道生成 80S 起始复合物。40S 小亚基之所以能在 AUG 处停下,可能是由于Met-tRNAi$^{Met}$ 的反密码子与 AUG 配对的结果。

起始过程中 mRNA 与 40S 小亚基结合时还需要 ATP,这可能是因为蛋白质合成中消除 mRNA 二级结构是一个耗能过程,需由 ATP 水解提供能量。另外根据“扫描模型”,在 40S 亚基沿 mRNA 移动过程中也需要能量。

### 4.4.3 肽链的延伸

生成起始复合物,第一个氨基酸(fMet/Met-tRNA)与核糖体结合以后,肽链开始伸长。按照 mRNA 模板密码子的排列,氨基酸通过新生肽键的方式(图 4-16)被有序地结合上去。肽链延伸由许多循环组成,每加一个氨基酸就是一个循环, 每个循环包括 AA-tRNA 与核糖体结合、肽键的生成和移位。

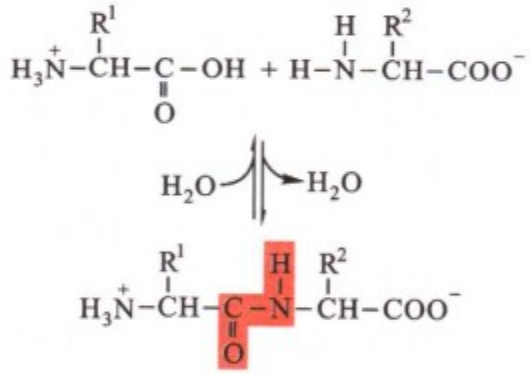


图 4-16 多肽链上肽键的形成——缩合反应

1. 后续 AA-tRNA 与核糖体结合

起始复合物形成以后,第二个 AA-tRNA 在延伸因子 EF-Tu 及 GTP 的作用下, 生成 AA-tRNA·EF-Tu·GTP 复合物,然后结合到核糖体的 A 位上。这时 GTP 被水解释放,通过延伸因子 EF-Ts 再生 GTP,形成 EF-Tu·GTP 复合物,进入新一轮循环(图 4-17)。

模板上的密码子决定了哪种 AA-tRNA 能被结合到 A 位上。由于 EF-Tu 只能与 fMet-tRNA 以外的其他 AA-tRNA 起反应, 所以起始 tRNA 不会被结合到 A 位上,这就是 mRNA 内部的 AUG 不会被起始 tRNA 读出,肽链中间不会出现甲酰甲硫氨酸的原因。

2. 肽键的生成

经过上一步反应后,在核糖体·mRNA·AA-tRNA 复合物中,AA-tRNA 占据 A 位,fMet-tRNA$^{fMet}$ 占据 P 位。在肽基转移酶(peptidyl transferase)的催化下,A 位上的 AA-tRNA 转移到 P 位,与 fMet-tRNA$^{fMet}$ 上的氨基酸生成肽键(图4-18)。起始 tRNA 在完成使命后离开核糖体 P 位点,A 位点准备接受新的 AA-tRNA,开

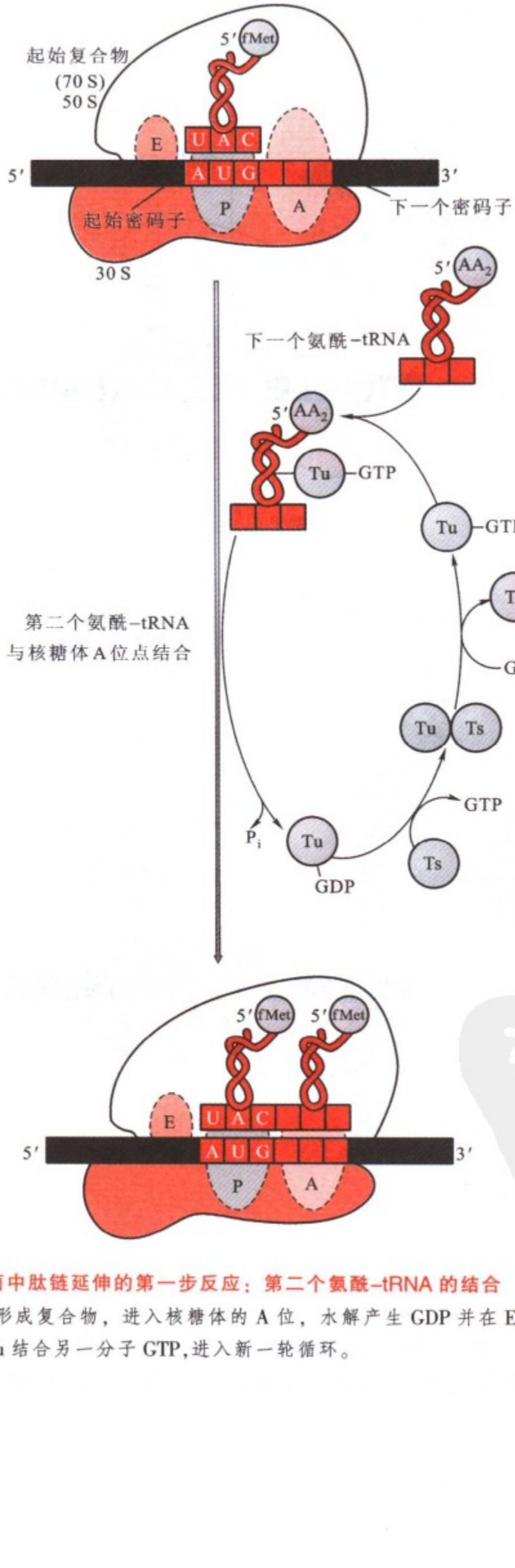



**图 4-17 细菌中肽链延伸的第一步反应：第二个氨酰-tRNA 的结合** 该氨酰-tRNA 首先与 EF-Tu·GTP 形成复合物，进入核糖体的 A 位，水解产生 GDP 并在 EF-Ts 的作用下释放 GDP 并使 EF-Tu 结合另一分子 GTP，进入新一轮循环。

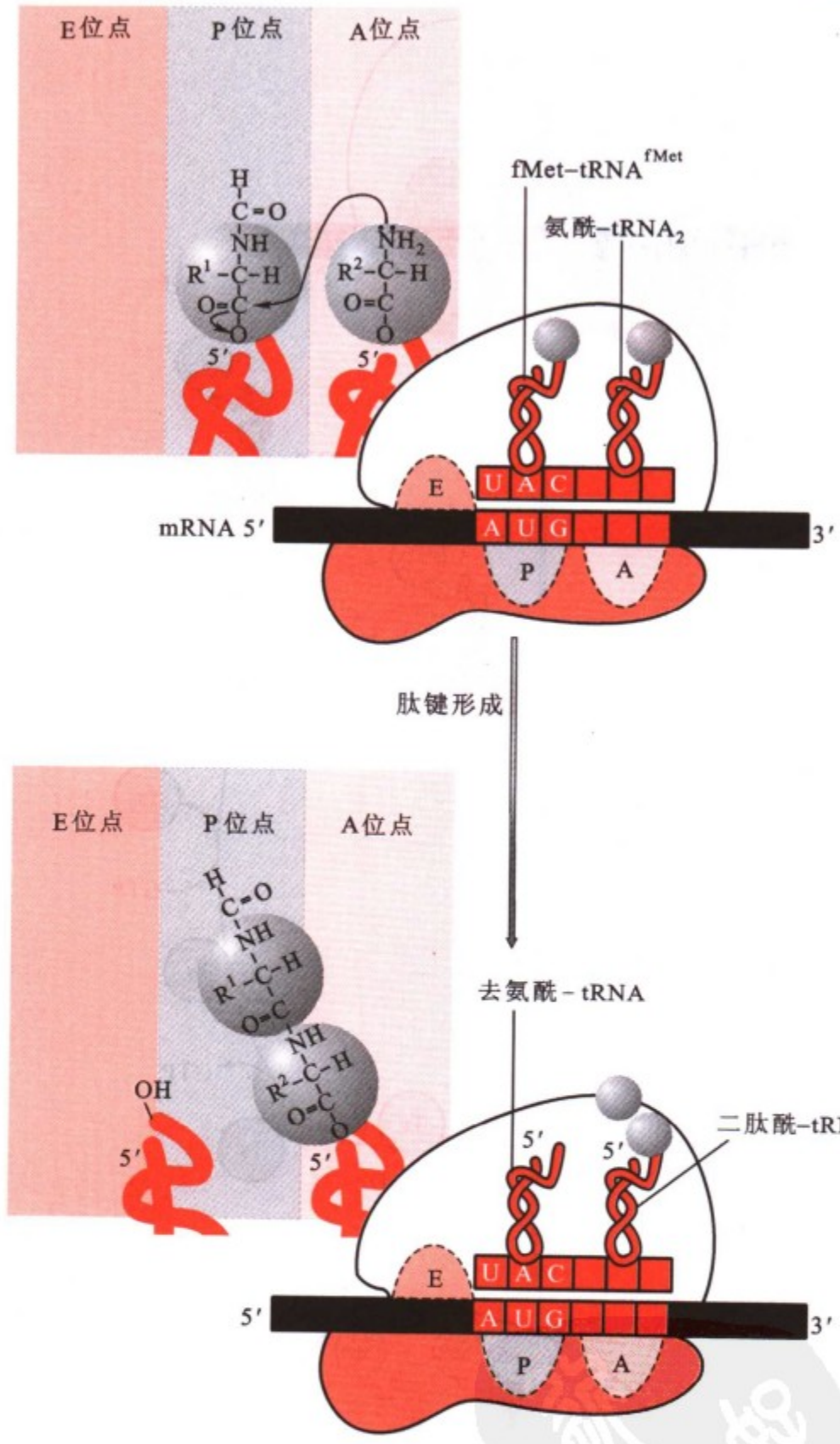



图 4-18 细菌中肽链延伸的第二步反应:肽链的生成

始下一轮合成反应。

3. 移位

肽键延伸过程中最后一步反应是移位,即核糖体向 mRNA 3′端方向移动一个密码子(图 4-19)。此时,仍与第二个密码子结合的二肽酰–tRNA$_2$ 从 A 位进入 P 位, 去氨酰–tRNA 被挤入 E 位,mRNA 上的第三位密码子则对应于 A 位。EF-G 是移位所必需的蛋白质因子,移位的能量来自另一分子 GTP 水解。

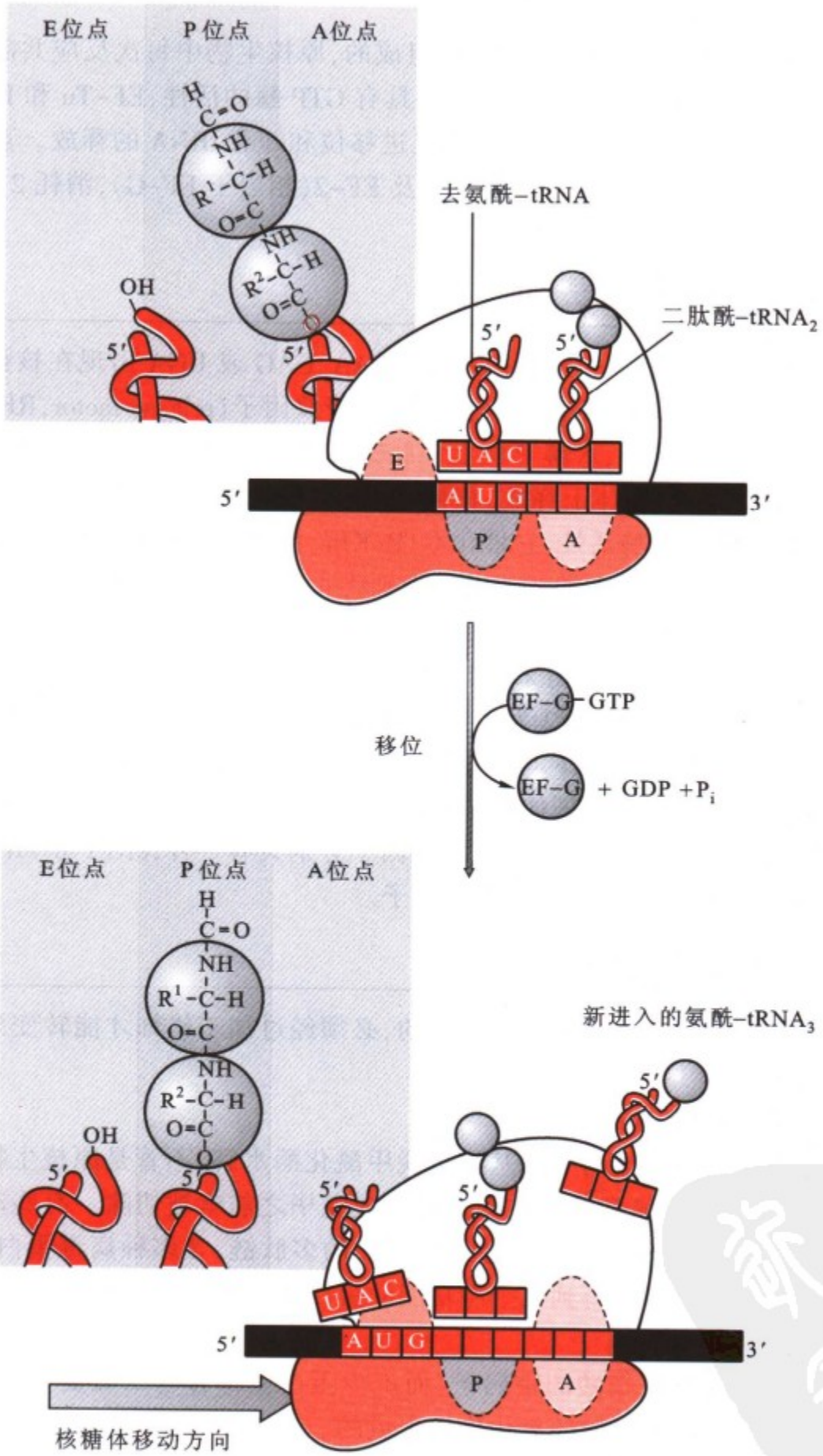



**图 4-19 细菌中肽链延伸的第三步反应:移位** 核糖体通过 EF-G 介导的 GTP 水解所提供的能量向 mRNA 模板 3′端移动一个密码子,使二肽酰-tRNA$_2$ 完全进入 P 位,准备开始新一轮肽链延伸。

用嘌呤霉素作为抑制剂做实验表明,核糖体沿 mRNA 移动与肽酰-tRNA 的移位这两个过程是偶联的。

肽链延伸是由许多个这样的反应组成的,原核生物中每次反应共需 3 个延伸因子,EF-Tu、EF-Ts 及 EF-G,它们都具有 GTP 酶的活性,EF-Tu 和 EF-Ts 能够促进 AA-tRNA 进入 A 位,EF-G 则促进移位和卸载 tRNA 的释放。真核生物细胞需 EF-1(对应于 EF-Tu 和 EF-Ts)及 EF-2(相当于 EF-G),消耗 2 个 GTP,向生长中的肽链加上一个氨基酸。

### 4.4.4 肽链的终止

肽链的延伸过程中,当终止密码子 UAA、UAG 或 UGA 出现在核糖体的 A 位时,没有相应的 AA-tRNA 能与之结合,而释放因子(release factor,RF)能识别这些密码子并与之结合,水解 P 位上多肽链与 tRNA 之间的二酯键。接着,新生的肽链和 tRNA 从核糖体上释放,核糖体大、小亚基解体,蛋白质合成结束。释放因子 RF 具有 GTP 酶活性,它催化 GTP 水解,使肽链与核糖体解离。

释放因子有两类。Ⅰ类释放因子识别终止密码子,并能催化新合成的多肽链从 P 位点的 tRNA 中水解释放出来;Ⅱ类释放因子在多肽链释放后刺激Ⅰ类释放因子从核糖体中解离出来。细菌细胞内存在三种不同的释放因子(RF1,RF2,RF3),其中,RF1 和 RF2 为Ⅰ类释放因子。RF1 能识别 UAG 和 UAA,RF2 识别 UGA 和 UAA。一旦 RF 与终止密码结合,它们就能诱导肽基转移酶把一个水分子而不是氨基酸加到延伸中的肽链上。RF3 为Ⅱ类释放因子,与核糖体的解离有关。真核细胞的Ⅰ类和Ⅱ类释放因子分别只有一种,eRF1 和 eRF3。Ⅰ类释放因子 eRF1 能够识别三个终止密码子。

### 4.4.5 蛋白质前体的加工

新生的多肽链大多数是没有功能的,必须经过加工修饰才能转变为有活性的蛋白质。

1. N 端 fMet 或 Met 的切除

细菌蛋白质氨基端的甲酰基能被脱甲酰化酶水解,不管是原核生物还是真核生物,N 端的甲硫氨酸往往在多肽链合成完毕之前就被切除。有些动物病毒如脊髓灰质炎病毒的 mRNA 可翻译成很长的多肽链,含多种病毒蛋白,经蛋白酶在特定位置上水解后得到几个有功能的蛋白质分子(图 4-20)。

2. 二硫键的形成

mRNA 中没有胱氨酸的密码子,而不少蛋白质都含有二硫键,这是蛋白质合成后通过两个半胱氨酸的氧化作用生成的。二硫键的正确形成对稳定蛋白的天然构象具有重要的作用。

3. 特定氨基酸的修饰

氨基酸侧链的修饰作用包括磷酸化(如核糖体蛋白质)、糖基化(如各种糖蛋白)、甲基化(如组蛋白、肌肉蛋白质)、乙基化(如组蛋白)、羟基化(如胶原蛋白)和羧基化等,图 4-21 是生物体内最普通发生修饰作用的氨基酸残基及其修

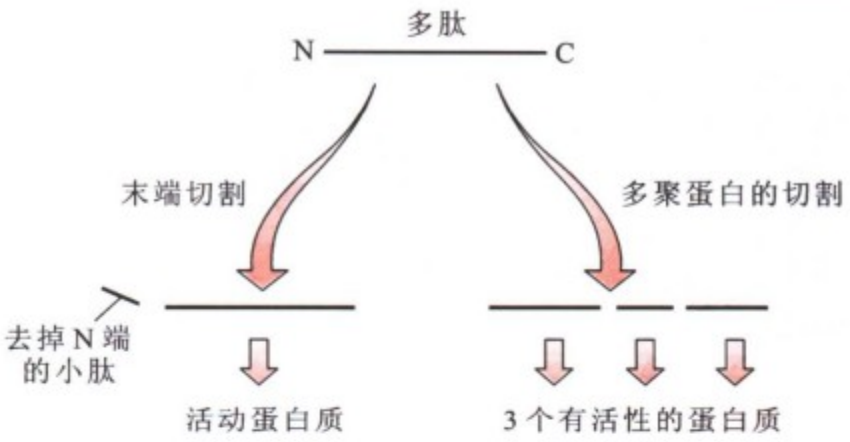



**图 4–20 新生蛋白质经蛋白酶切割后变成有功能的成熟蛋白质** 左：新生蛋白质在去掉 N 端一部分残基后变成有功能的蛋白质；右：某些病毒或细菌可合成无活性的多聚蛋白质，经蛋白酶切割后成为有功能的成熟蛋白。

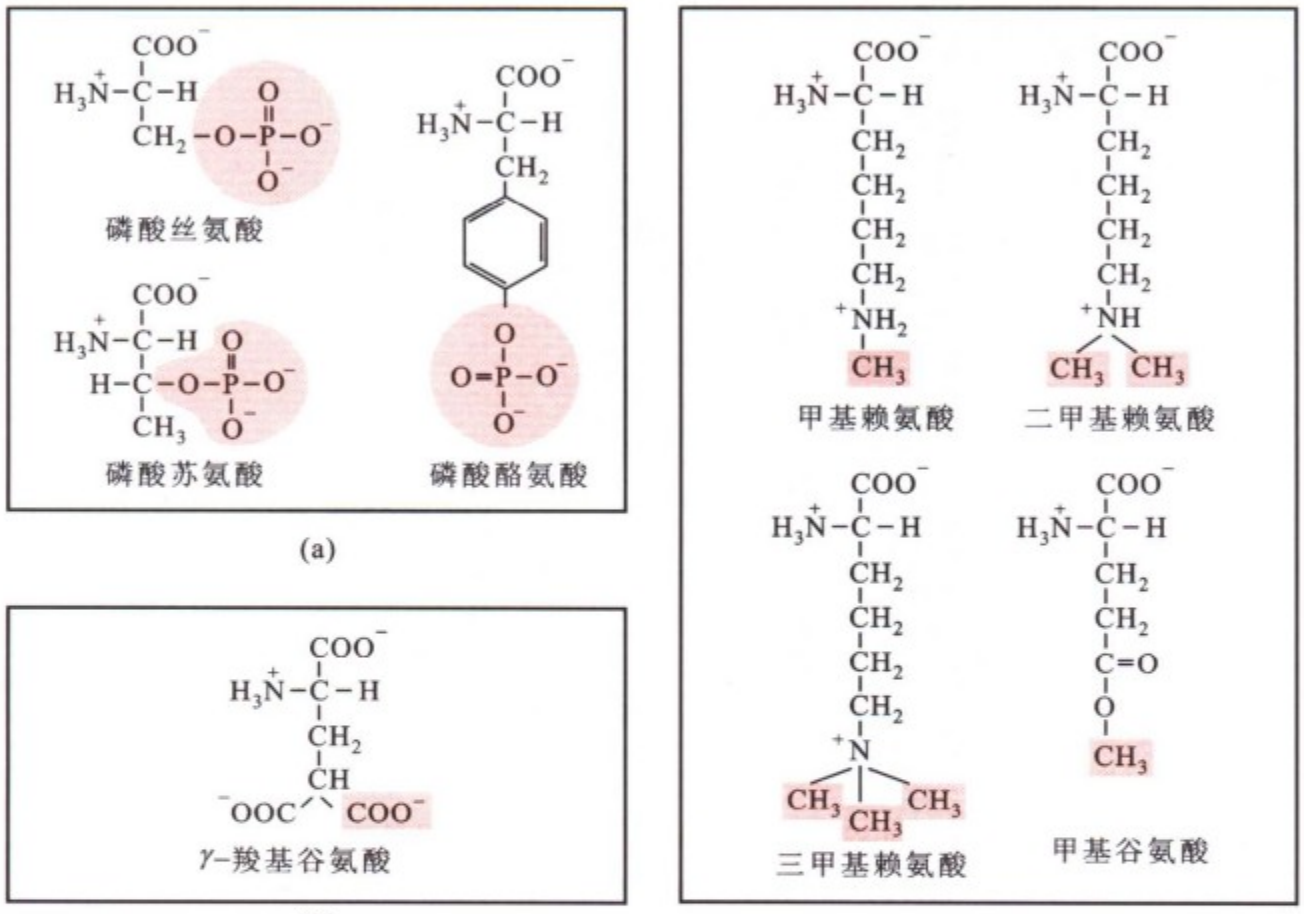



**图 4–21 生物体内最常见的被修饰的氨基酸及其修饰产物**

饰产物，图 4–22 是组蛋白 H3(共有 135 个氨基酸残基)N 端 35 个残基中可能出现的 4 种修饰。糖蛋白主要是通过蛋白质侧链上的天冬氨酸、丝氨酸、苏氨酸残基加上糖基形成的，胶原蛋白上的脯氨酸和赖氨酸多数是羟基化的。实验证明，内质网可能是蛋白质 *N*–糖基化的主要场所(图 4–23)。

(1) 磷酸化 主要由多种蛋白激酶催化，发生在丝氨酸、苏氨酸和酪氨酸等三种氨基酸的侧链。

**图 4–22 发生在小牛组蛋白 H3 前 30 个氨基酸残基中的 4 种化学修饰**

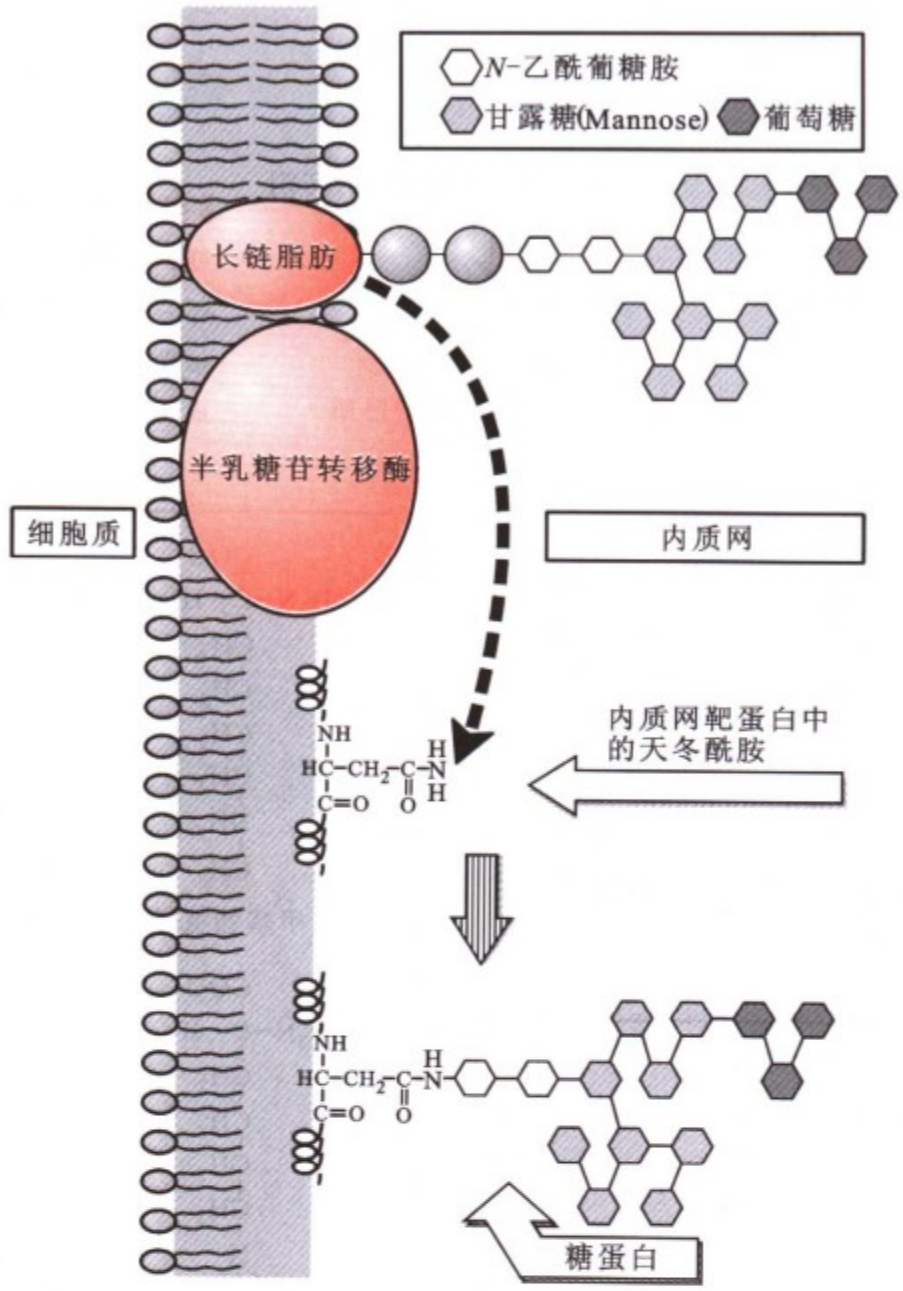


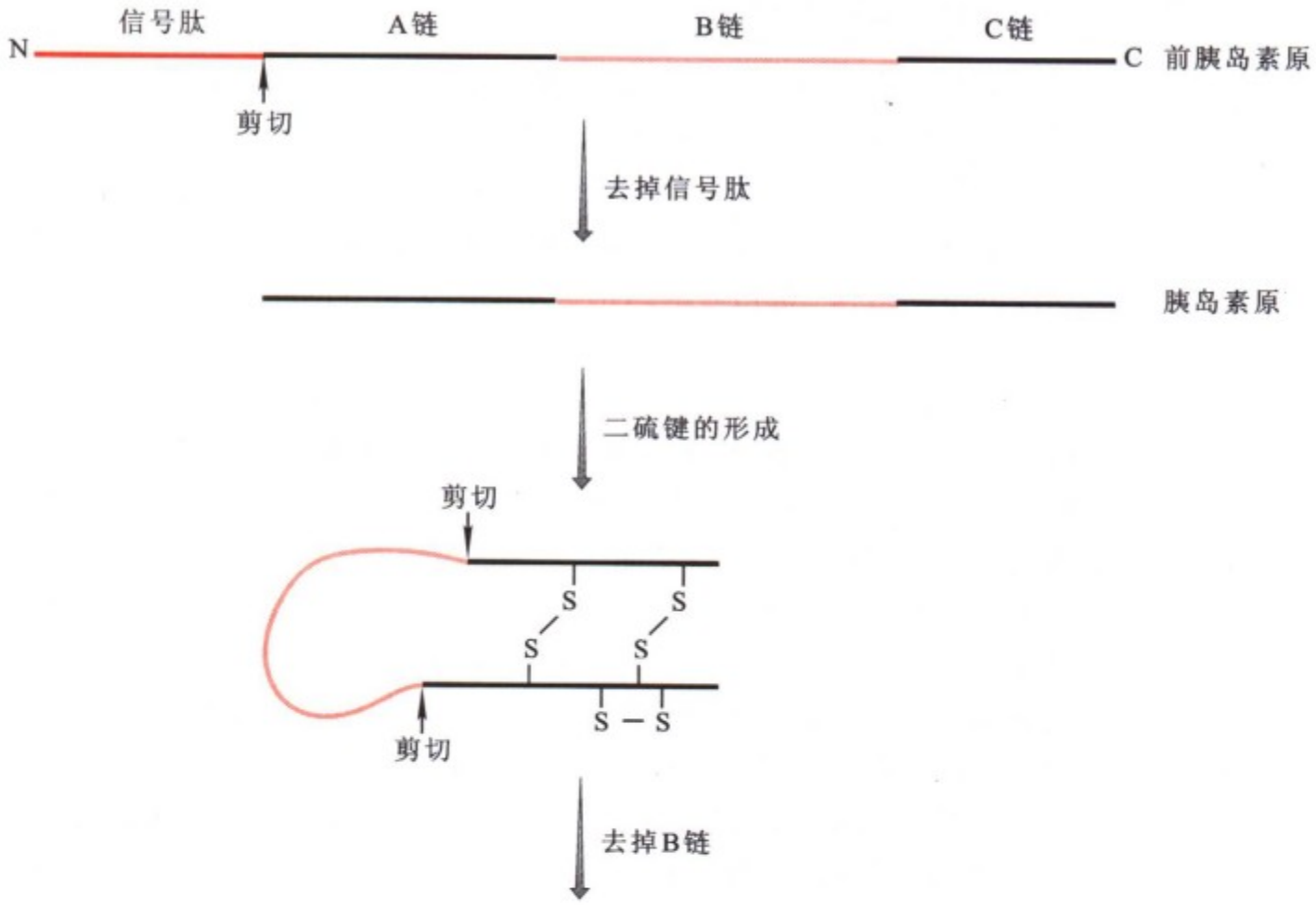

图 4-23 蛋白质通过其 N 端的信号肽在内质网中运转到不同的细胞器

(2) 糖基化 它是真核细胞蛋白质的特征之一。大多数糖基化是由内质网中的糖基化酶(glycosylase)催化进行的。所有的分泌蛋白和膜蛋白几乎都是糖基化蛋白质。

(3) 甲基化 蛋白质的甲基化是由 *N*-甲基转移酶催化的,该酶主要存在于细胞质内。甲基化包括发生在 Arg、His 和 Gln 的侧基的 *N*-甲基化以及 Glu 和 Asp 侧基的 *O*-甲基化。

4. 切除新生肽链中的非功能片段

如新合成的胰岛素前体是前胰岛素原,必须先切去信号肽变成胰岛素原,再切去 C-肽,才变成有活性的胰岛素(图 4-24)。不少多肽类激素和酶的前体都要经过加工才能变为活性分子,如血纤维蛋白原、胰蛋白酶原经过加工切去部分肽段才能成为有活性的血纤维蛋白、胰蛋白酶。蜂毒素能溶解动物细胞,也能溶解蜜蜂自身的细胞,所以只能在细胞内合成没有活性的前毒素,分泌进入刺吸器后,其 N 端的 22 个氨基酸残基被蛋白酶水解,才能生成有毒性的功能蛋白质(图 4-25)。一般



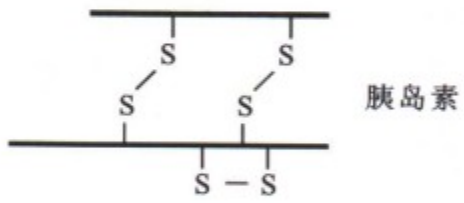




**图 4-24 前胰岛素原蛋白翻译后成熟过程示意图**

**图 4-25 蜂毒蛋白只有经蛋白酶水解切除 N 端的 22 个氨基酸以后才具有生物活性** 该胞外蛋白酶只能特异性切割 X-Y2 肽。其中 X 是丙氨酸,天冬氨酸和谷氨酸;Y 是丙氨酸或脯氨酸。

说来,由多个肽链及其他辅助成分构成的蛋白质,在多肽链合成后还需经过多肽链之间以及多肽链与辅基之间的聚合过程,才能成为有活性的蛋白质。

### 4.4.6 蛋白质的折叠

由核糖体合成的所有新生肽链必须通过正确的折叠才能形成动力学和热力学稳定的三维构象,从而表现出生物学活性或功能。因此,可以说蛋白质折叠是翻译后形成功能蛋白质的必经阶段。如果蛋白质折叠错误,其生物学功能就会受到影响或丧失,严重者甚至会引起疾病。新合成的蛋白质分子如何形成具有功能的空间结构?蛋白质的结构与功能的关系等问题已成为结构生物学研究的热点。

多肽链的折叠是一个复杂的过程,新生多肽一般首先折叠成二级结构,然后再进一步折叠盘绕成三级结构。对于单链多肽蛋白质,三级结构就已具有蛋白质的功能;对于寡聚蛋白质,一般需要进一步组装成更为复杂的四级结构,才能表现出天然蛋白的活性或功能。

有些蛋白质只有在另一些蛋白质存在的情况下才能正确完成折叠过程,形成功能蛋白质。分子伴侣(molecular chaperone)是目前研究得比较多的能够在细胞内辅助新生肽链正确折叠的蛋白质。它是一类序列上没有相关性但有共同功能的保守性蛋白质,它们在细胞内能帮助其他多肽进行正确的折叠、组装、运转和降解。目前认为细胞内至少有两类分子伴侣家族,即热休克蛋白(heat shock protein)家族和伴侣素(chaperonin)。

1. 热休克蛋白

它是一类应激反应性蛋白,包括 HSP70、HSP40 和 GrpE 三个家族,广泛存在于原核及真核细胞中。三者协同作用,促使某些能自发折叠的蛋白质正确折叠形成天然空间构象。

2. 伴侣素

伴侣素包括 HSP60 和 HSP10(原核细胞中的同源物分别为 GroEL 和 GroES),它主要是为非自发性折叠蛋白提供能折叠形成天然结构的微环境。

伴侣分子在新生肽链折叠中主要通过防止或消除肽链的错误折叠,增加功能性蛋白质折叠产率来发挥作用,而并非加快折叠反应速度。分子伴侣本身并不参与最终产物的形成。

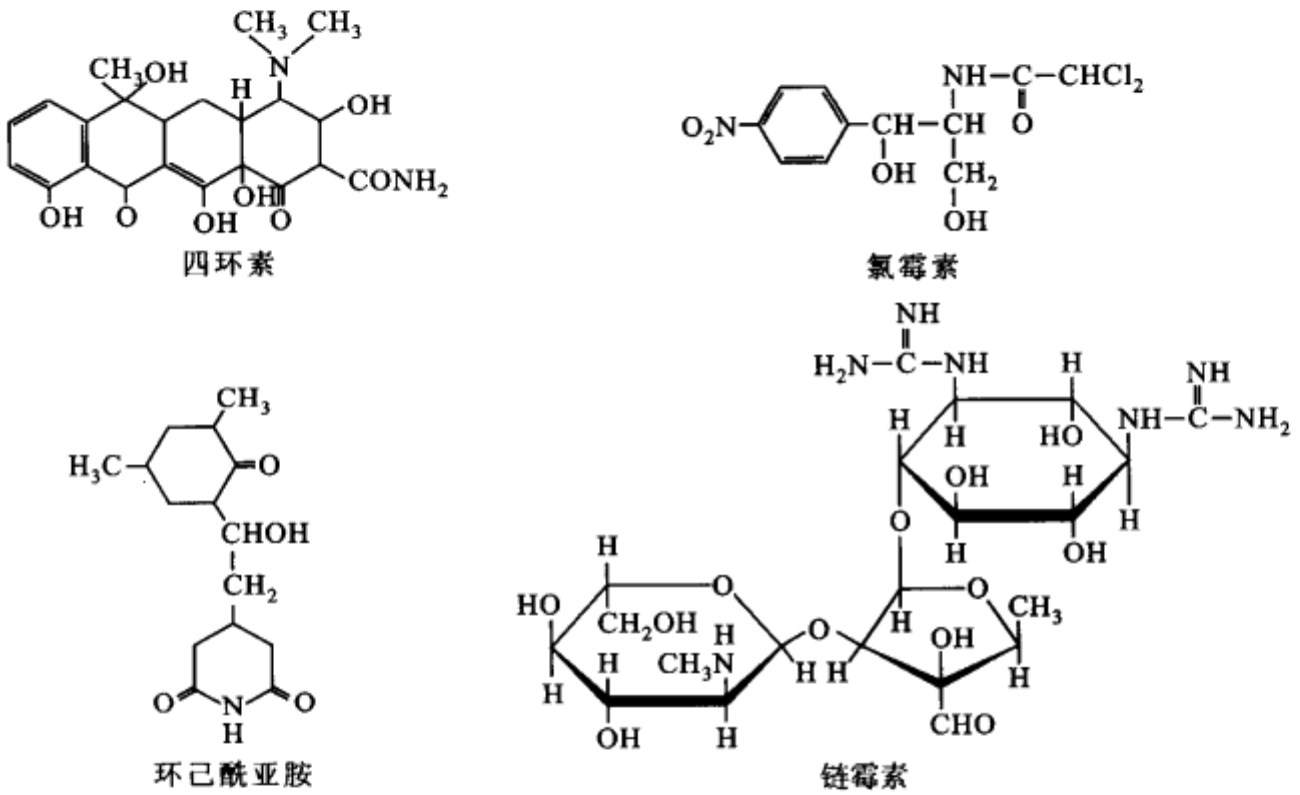
### 4.4.7 蛋白质合成的抑制剂

蛋白质生物合成的抑制剂主要是一些抗生素,如嘌呤霉素、链霉素、四环素、氯霉素、红霉素等,此外,如 5-甲基色氨酸、环己亚胺、白喉毒素、干扰素、蓖麻蛋白和其他核糖体灭活蛋白等都能抑制蛋白质的合成。这些抑制剂不仅对于研究蛋白质的合成机制十分重要,也是在临床上治疗细菌感染的重要药物。

抗生素对蛋白质合成的作用可能是阻止 mRNA 与核糖体结合(氯霉素),

或阻止 AA-tRNA 与核糖体结合(四环素类),或干扰 AA-tRNA 与核糖体结合而产生错读(链霉素、新霉素、卡那霉素等),或作为竞争性抑制剂抑制蛋白质合成。

链霉素是一种碱性三糖,可以多种方式抑制原核生物核糖体,能干扰 fMet-tRNA 与核糖体的结合,从而阻止蛋白质合成的正确起始,也会导致 mRNA 的错读。若以 poly U 作模板,则除苯丙氨酸(UUU)外,异亮氨酸(AUU)也会被掺入。用抗链霉素细菌的 50S 亚基及对链霉素敏感细菌的 30S 亚基重组获得的核糖体对链霉素是敏感的,而用敏感菌的 50S 亚基及抗性菌的 30S 亚基组成的核糖体对链霉素有抗性,表明链霉素的作用位点在 30S 亚基上。图 4-26 是几种常见蛋白质合成抑制剂的化学结构式。




▌图 4-26 几种常见蛋白质合成抑制剂的结构式

嘌呤霉素是 AA-tRNA 的结构类似物，不需要延伸因子就可以结合在核糖体的 A 位上，抑制 AA-tRNA 的进入。它所带的氨基与 AA-tRNA 上的氨基一样,能与生长中的肽链上的羧基反应生成肽键,这个反应的产物是一条 3′羧基端挂了一个嘌呤霉素残基的小肽，肽酰嘌呤霉素随后从核糖体上解离出来,所以嘌呤霉素是通过提前释放肽链来抑制蛋白质合成的(图 4-27)。

此外,青霉素、四环素和红霉素只与原核细胞核糖体发生作用,从而阻遏原核生物蛋白质的合成,抑制细菌生长。氯霉素和嘌呤霉素既能与原核细胞核糖体结合,又能与真核生物核糖体结合,妨碍细胞内蛋白质合成,影响细胞生长。因此,前 3 种抗生素被广泛用于人类医学,后两种则很少在医学上使用。氯霉素虽然有时也用于治病,但医生只在前几种抗生素均不见效时才使用,剂量和周期也控制得较严。

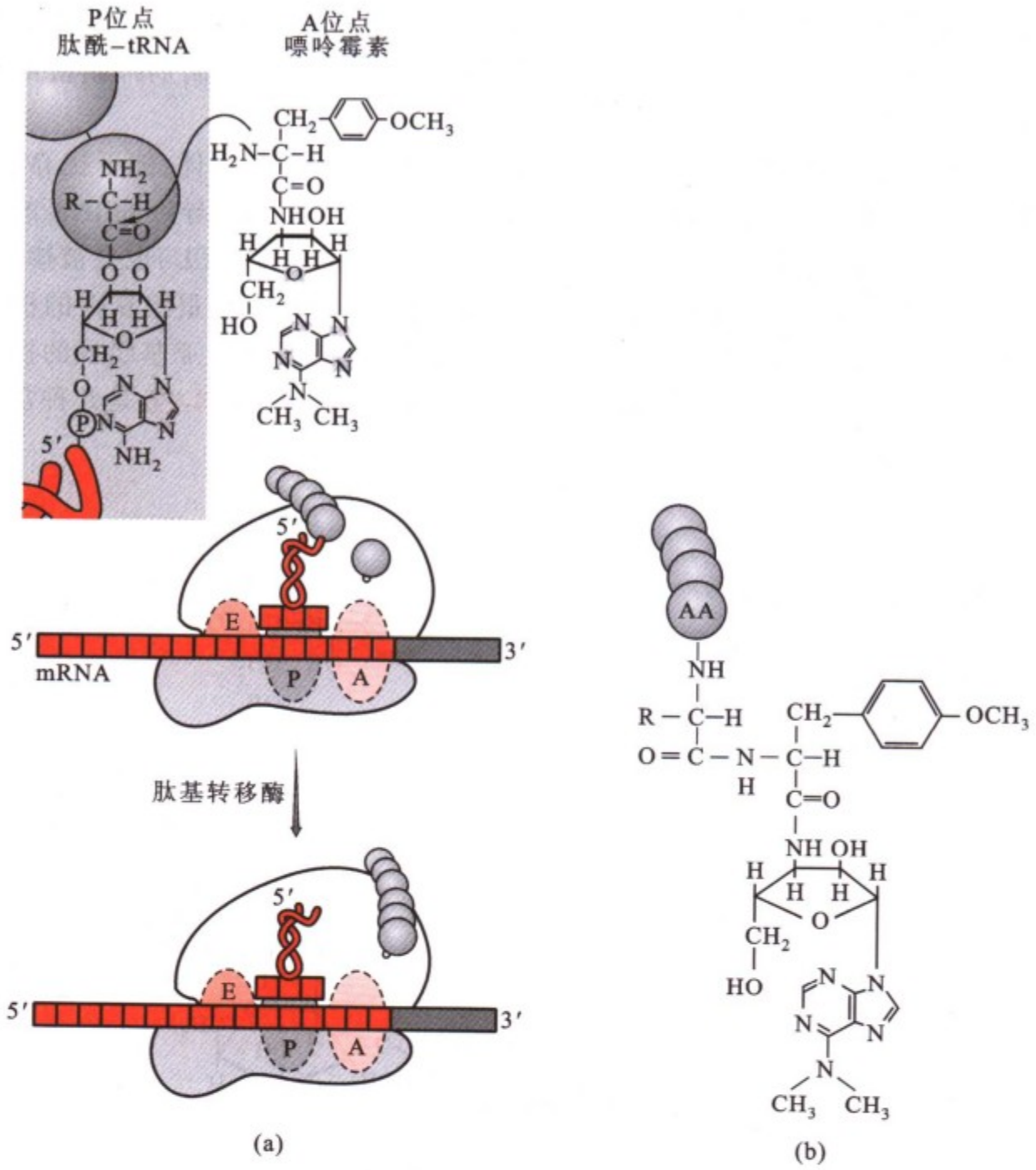



图 4-27 嘌呤霉素抑制蛋白质合成的分子机制 (a)嘌呤霉素的结构类似于带有氨酰的 tRNA,能与核糖体 A 位点相结合;(b)肽基嘌呤霉素。

## 4.5 蛋白质转运机制

在生物体内,蛋白质的合成位点与功能位点常常被一层或多层细胞膜所隔开,这样就产生了蛋白质转运的问题。核糖体是真核生物细胞内合成蛋白质的场所,几乎在任何时候,都有数以百计或千计的蛋白质离开核糖体并被输送到细胞质、细胞核、线粒体、内质网和溶酶体、叶绿体等各个部分,补充和更新细胞功能。由于细胞各部分都有特定的蛋白质组分,因此合成的蛋白质必须准确无误地定向运送才能保证生命活动的正常进行。蛋白质是怎样从合成部位运送至功能部位的?它们又是如何跨膜运送的?跨膜之后又是依靠什么信息到达各自"岗位"的?对于膜蛋白来说,究竟是什么因素决定它为外周蛋白还是内在蛋白?

是部分镶嵌还是跨膜分布？在膜的外侧还是内侧？这些都是十分有趣的问题，也是生物膜研究中非常活跃的领域(图 4-28)。

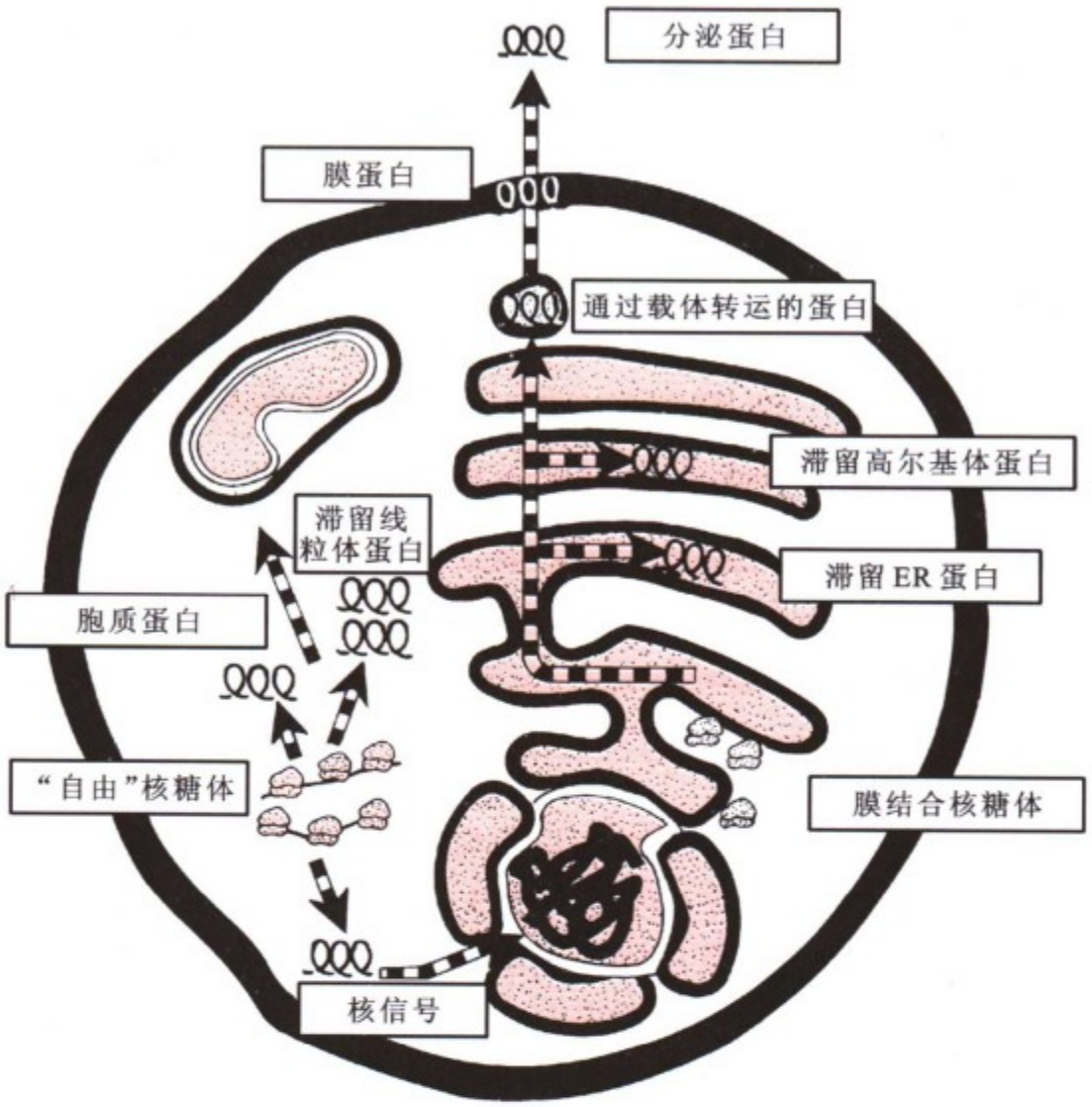



图 4-28 蛋白质合成和转运过程示意图

一般说来，蛋白质转运可分为两大类：若某个蛋白质的合成和转运是同时发生的，则属于翻译-转运同步机制；若蛋白质从核糖体上释放后才发生转运，则属于翻译后转运机制。这两种转运方式都涉及蛋白质分子内特定区域与细胞膜结构的相互关系。

表 4-14 列举了跨膜运输和镶入膜内的几种主要蛋白质。从表中可知，分泌蛋白质大多是以翻译-转运同步机制运输的。在细胞器发育过程中，由细胞质进入细胞器的蛋白质大多是以翻译后转运机制运输的。而参与生物膜形成的蛋白质，则依赖于上述两种不同的转运机制镶入膜内。

表 4-14 几类主要蛋白质的转运机制

| 蛋白性质 | 转运机制 | 主要类型 |
| --- | --- | --- |
| 分泌 | 蛋白质在结合核糖体上合成，并以翻译转运同步机制运输 | 免疫球蛋白、卵蛋白、水解酶、激素等 |
| 细胞器发育 | 蛋白质在游离核糖体上合成，以翻译后转运机制运输 | 核、叶绿体、线粒体、乙醛酸循环体、过氧化物酶体等细胞器中的蛋白质 |
| 膜的形成 | 两种机制兼有 | 质膜、内织网、类囊体中的蛋白质 |

### 4.5.1 翻译–转运同步机制

一般认为,蛋白质定位的信息存在于该蛋白质自身结构中,并且通过与膜上特殊受体的相互作用得以表达,这就是信号肽假说的基础。这一假说认为,蛋白质跨膜转运信号也是由 mRNA 编码的。在起始密码子后,有一段编码疏水性氨基酸序列的 RNA 区域,被称为信号序列(图 4–29)。信号序列在结合核糖体上合成后便与膜上特定受体相互作用,产生通道,允许这段多肽在延长的同时穿过膜结构,因此,这种方式是边翻译边跨膜转运。

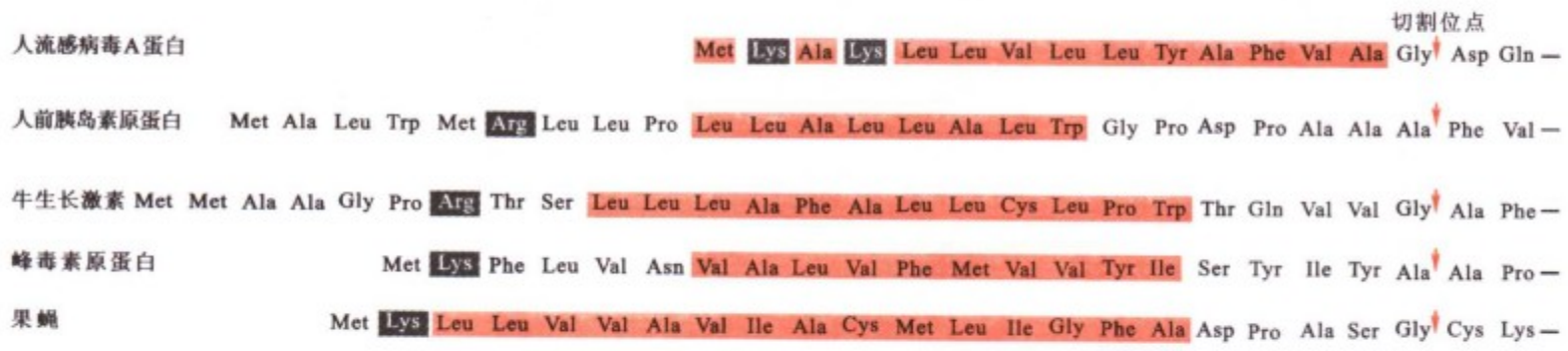



图 4–29 蛋白质通过其 N 端的信号肽在内质网中转运到不同的细胞器

绝大部分被运入内质网内腔的蛋白质都带有一个信号肽,该序列常常位于蛋白质的氨基端,长度一般在 13~36 个残基,有如下三个特点:①一般带有10~15 个疏水氨基酸;②在靠近该序列 N 端常常有 1 个或数个带正电荷的氨基酸;③在其 C 端靠近蛋白酶切割位点处常常带有数个极性氨基酸,离切割位点最近的那个氨基酸往往带有很短的侧链(丙氨酸或甘氨酸)。

根据信号肽假说,同细胞质中其他蛋白质的合成一样,分泌蛋白的生物合成开始于结合核糖体,当翻译进行到大约 50~70 个氨基酸残基之后,信号肽开始从核糖体的大亚基露出,被粗糙内质网膜上的受体识别,并与之结合。信号肽过膜后被内质网腔的信号肽酶水解,正在合成的新生肽随之通过蛋白孔道穿越疏水的双层磷脂。一旦核糖体移到 mRNA 的“终止”密码子,蛋白质合成即告完成,翻译体系解散,膜上的蛋白孔道消失,核糖体重新处于自由状态(图 4–30)。

如果说信号序列在决定蛋白质的特定去向上起着指导作用,那么,信号肽在蛋白质运输过程中是如何起作用的呢?已知野生型细胞中,β–半乳糖酶定位于胞质内,麦芽糖结合蛋白定位于胞外,麦芽糖转运蛋白位于外膜内腔。如果把 β–半乳糖酶羧基端基因片段与麦芽糖转运蛋白和麦芽糖结合蛋白的氨基端基因片段相连,并以此为模板指导合成杂种蛋白质,就可以通过测定 β–半乳糖酶的活性确定杂种蛋白质在细胞内的位置。改变融合基因中编码麦芽糖转运蛋白氨基端的基因序列,得到了如下结论:

① 完整的信号多肽是保证蛋白质转运的必要条件。信号序列中疏水性氨基酸突变成亲水性氨基酸后,会阻止蛋白质转运而使新生蛋白质以前体形式积累在胞质中。

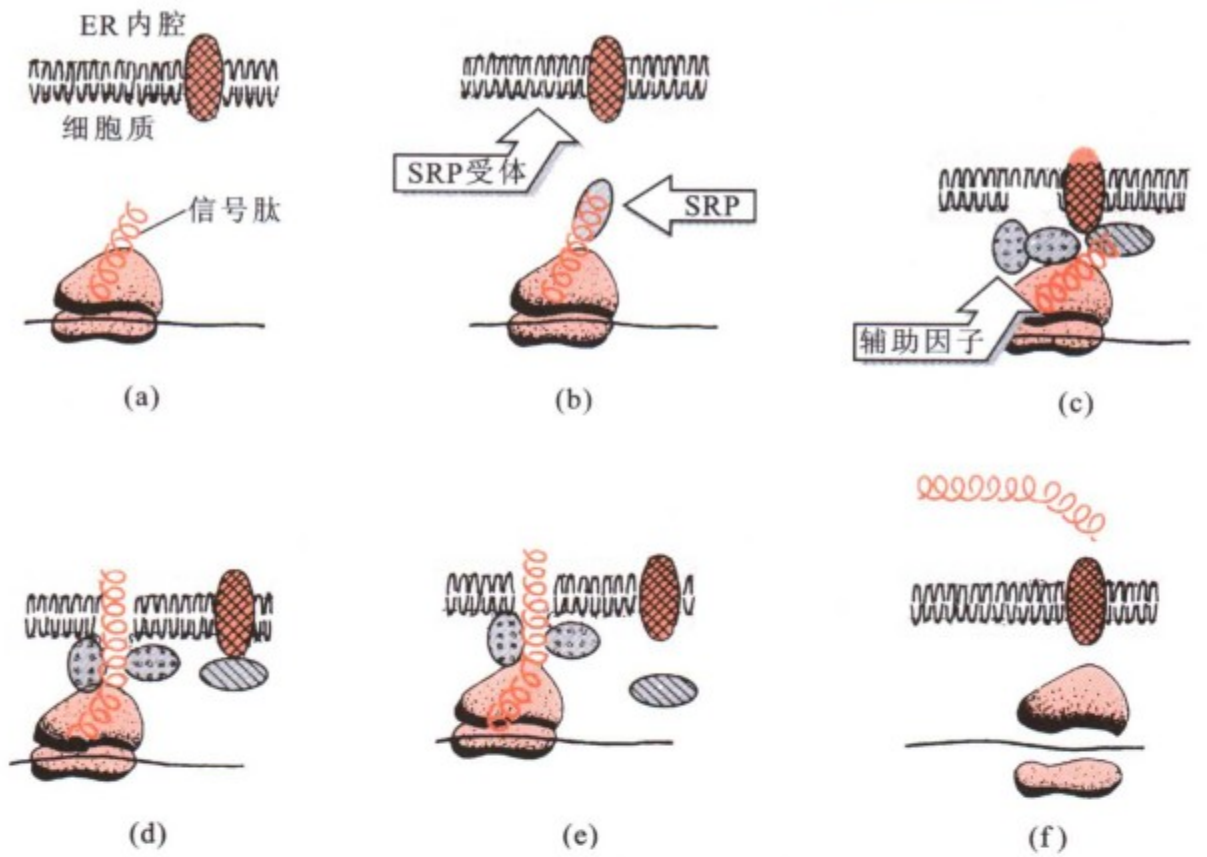



**图 4-30 蛋白质跨膜转运的信号肽假说及其运输过程** (a)蛋白质合成起始首先合成信号肽;(b)SRP与信号肽结合,翻译暂停;(c)SRP与SRP受体相结合,核糖体与膜结合,翻译重新开始;(d)信号肽进入膜结构;(e)蛋白质过膜,信号肽被切除,翻译继续进行;(f)蛋白质完全过膜,核糖体解离。

② 仅有信号肽还不足以保证蛋白质转运的发生。把麦芽糖转运蛋白中完整的、可水解的信号序列加到β-半乳糖酶上,发现杂种蛋白质仍以前体形式留在胞质内。所以说,要使蛋白质顺利跨膜,还要求转运蛋白质在信号序列以外的部分有相应的结构变化。因此,科学家认为,指导蛋白质转运的信号事实上可能要长于被蛋白酶降解掉的信号肽链。

③ 信号序列的切除并不是运转所必需的。如果把细菌外膜脂蛋白信号序列中的甘氨酸残基突变成天冬氨酸残基,便能抑制该蛋白信号肽的水解,但不能抑制其跨膜转运。

④ 并非所有的转运蛋白质都有可降解的信号肽。卵清蛋白是以翻译-转运同步机制进入微粒体中的,但它并没有可降解的信号序列。有人认为,在卵清蛋白分子的中心区域有相当于信号肽功能的肽段存在。据此,"信号肽"应当定义为:能启动蛋白质转运的任何一段多肽。

研究新生肽向内质网内腔转运过程发现,SRP(信号识别蛋白)和DP(停靠蛋白,又称SRP受体蛋白)介导了蛋白质的跨膜转运过程。SRP能同时识别正在合成需要通过内质网膜进行转运的新生肽和自由核糖体,它与这类核糖体上新生蛋白的信号肽结合是多肽正确转运的前提,但同时也导致了该多肽合成的暂时终止(此时新生肽一般长约70个残基左右)。SRP-信号肽-多核糖体复合物即被引向内质网膜并与SRP的受体——DP相结合。只有当SRP与DP相结合时,多肽合成才恢复进行,信号肽部分通过膜上的核糖体受体及蛋白转运复合

物跨膜进入内质网内腔,新生肽链重新开始延伸。整个蛋白跨膜以后,信号肽被水解,形成高级结构和成熟型蛋白质,并被运送到相应细胞器。SRP 与 DP 的结合很可能导致受体聚集而形成膜孔道,使信号肽及与其相连的新生肽得以通过。此时,SRP 与 DP 相分离并恢复游离状态。待翻译过程结束后,核糖体的大、小亚基解离,受体解聚,通道消失,内质网膜也恢复完整的脂双层结构(图 4-31)。进入内质网内腔后,蛋白质常以转运载体的形式被送入高尔基体或形成转运小泡,分别运送到各自的亚细胞位点。

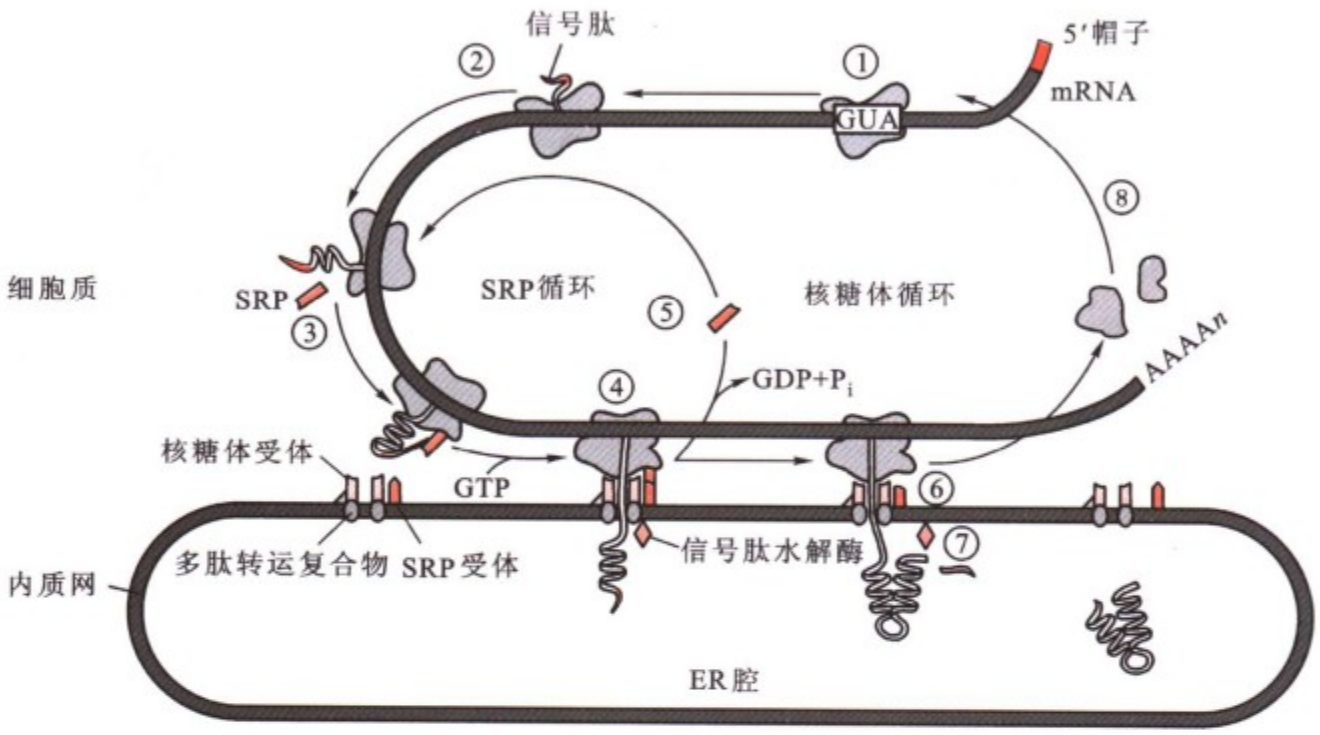



**图 4-31 新生蛋白质通过同步转运途径进入内质网内腔的主要过程** ①核糖体组装、翻译起始;②位于蛋白质 N 端的信号肽序列首先被翻译;③SRP 与核糖体、GTP 以及带有信号肽的新生蛋白质相结合,暂时中止肽链延伸;④核糖体-SRP 复合物与膜上的受体相结合;⑤GTP 水解,释放 SRP 并进入新一轮循环;⑥肽链重新开始延伸并不断向内腔运输;⑦信号肽被切除;⑧多肽合成结束,核糖体解离并恢复到翻译起始前的状态。

## 4.5.2 翻译后转运机制

研究发现,叶绿体和线粒体中有许多蛋白质和酶是由细胞质提供的,其中绝大多数以翻译后转运机制进入细胞器内。

1. 线粒体蛋白质跨膜转运

线粒体是细胞的“动力站”,它虽然含有遗传物质(DNA、RNA)以及核糖体等,但它的 DNA 信息含量有限,大部分线粒体蛋白质都是由核 DNA 编码,在细胞质自由核糖体上合成。被释放至细胞质,再跨膜转运到线粒体各部分。与分泌蛋白质通过内质网膜进行转运不同,通过线粒体膜的蛋白质是在合成以后再转运的。这个过程有如下特征:

① 通过线粒体膜的蛋白质在转运之前大多数以前体形式存在,它由成熟蛋白质和位于 N 端的一段前导肽(leader peptide)共同组成。迄今已有

40 多种线粒体蛋白质前导肽的一级结构被阐明，它们约含 20~80 个氨基酸残基，当前体蛋白跨膜时，前导肽被一种或两种多肽酶所水解，释放成熟蛋白质。

② 蛋白质通过线粒体内膜的转运是一种需要能量的过程。

③ 蛋白质通过线粒体膜转运时，首先由外膜上的 Tom 受体复合蛋白识别与 Hsp70 或 MSF 等分子伴侣相结合的待转运多肽，通过 Tom 和 Tim 组成的膜通道进入线粒体内腔。蛋白质跨膜转运时的能量来自线粒体 Hsp70 引发的 ATP 水解和膜电位差(图 4-32)。

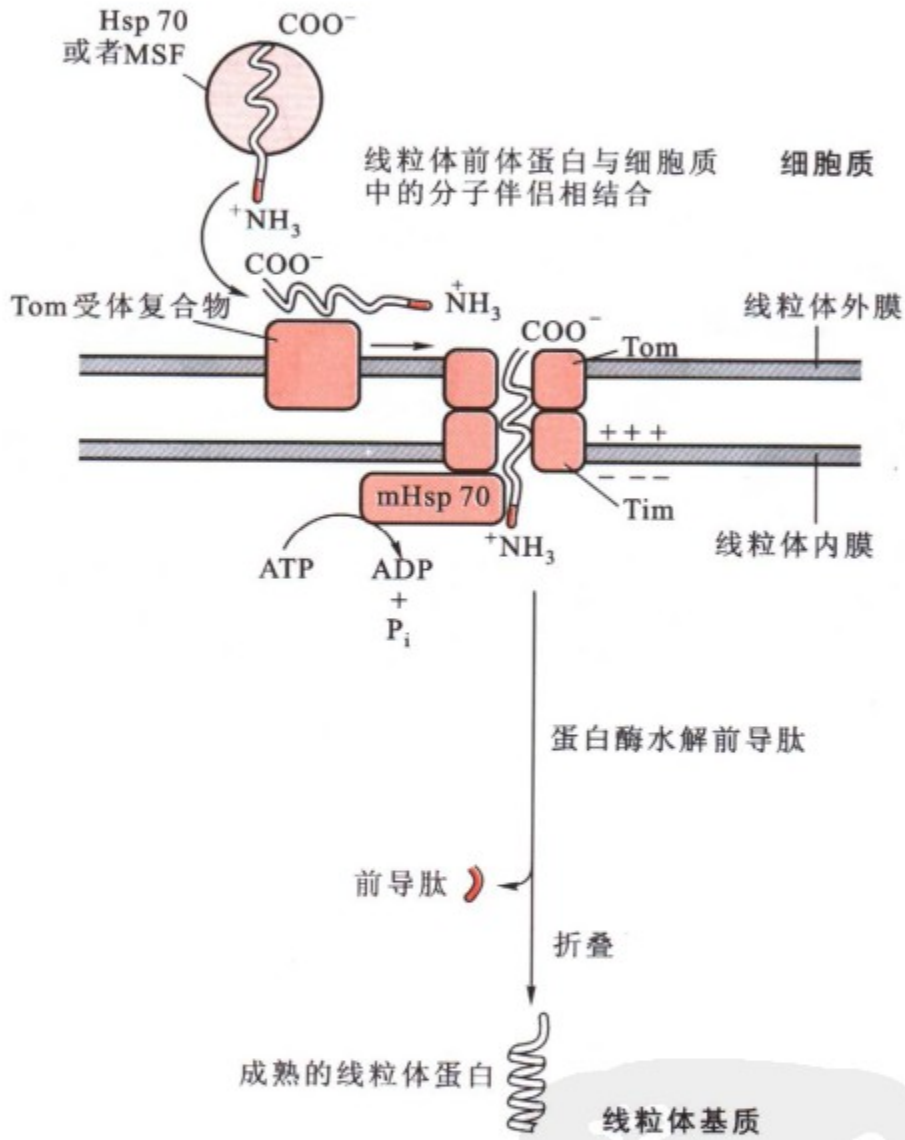



图 4-32 线粒体蛋白质跨膜转运

2. 前导肽的作用与性质

拥有前导肽的线粒体蛋白质前体能够跨膜转运进入线粒体，在这一过程中前导肽被水解，前体转变为成熟蛋白，失去继续跨膜能力。因此，前导肽对线粒体蛋白质的识别和跨膜转运显然起着关键作用。

前导肽一般具有如下特性：带正电荷的碱性氨基酸(特别是精氨酸)含量较为丰富，它们分散于不带电荷的氨基酸序列之间；缺少带负电荷的酸性氨基酸；羟基氨基酸(特别是丝氨酸)含量较高；有形成两亲(既有亲水又有疏水部分)α 螺旋结构的能力。带正电荷的碱性氨基酸在前导肽中有重要的作用，如果它们

被不带电荷的氨基酸所取代,就不能发挥牵引蛋白质跨膜的作用。

前导肽跨膜转运时首先与线粒体外膜上的受体相结合,实验表明,并非线粒体蛋白质前体都共用一种受体,但 Tom 受体可能是线粒体蛋白质跨膜转运时最主要的受体蛋白。线粒体有内、外两层膜,前导肽的不同部位可能在蛋白的跨膜运输过程中发挥不同的作用(图 4-33)。有些前导肽含有“止运入”肽段,当该肽段被跨膜通道中的受体蛋白识别时,所运输的多肽将被定位在膜上。

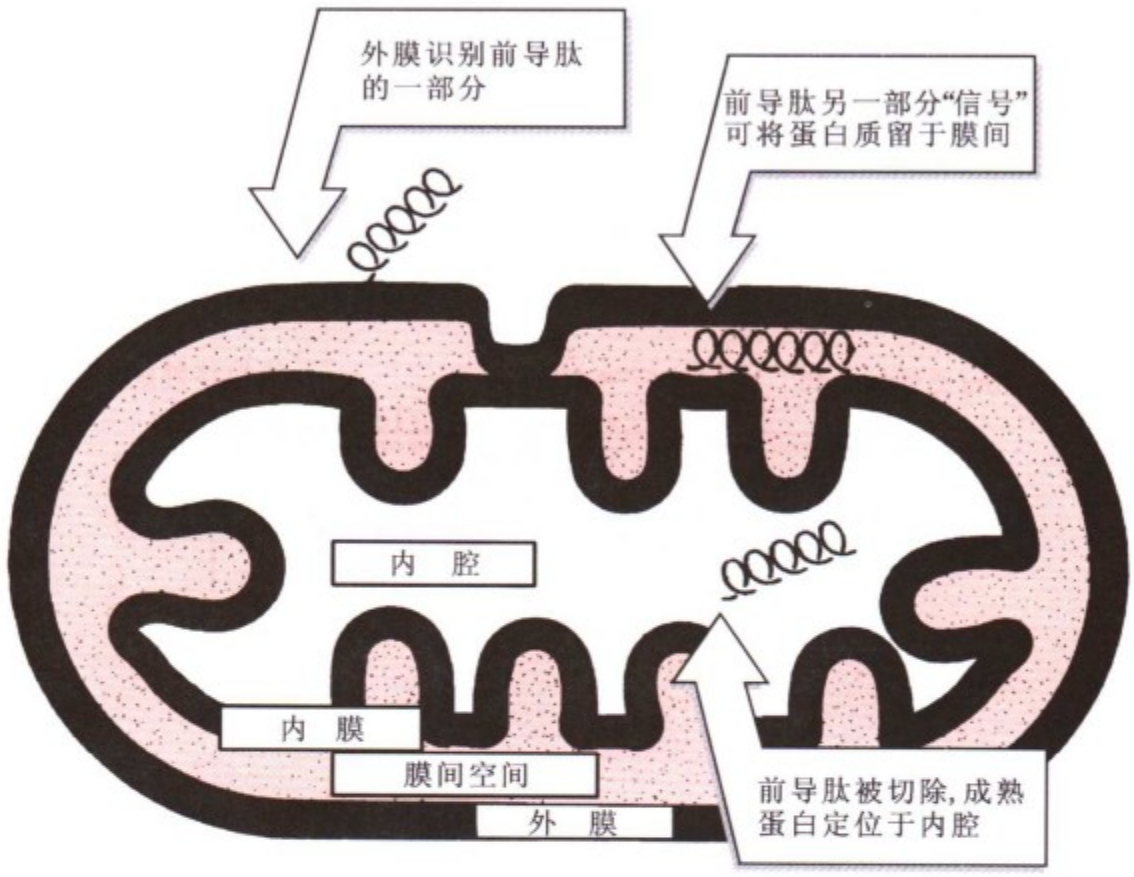



图 4-33 前导肽的不同区域可能在蛋白质跨膜转运过程中起不同的作用

3. 叶绿体蛋白质的跨膜转运

大多数科学家认为,叶绿体多肽是胞质中的游离核糖体上合成后脱离核糖体并折叠成具有三级结构的蛋白质分子,多肽上某些特定位点结合于只有叶绿体膜上才有的特异受体位点。叶绿体定位信号肽一般有两个部分,第一部分决定该蛋白质能否进入叶绿体基质,第二部分决定该蛋白能否进入类囊体(图 4-34)。在这一模型中,蛋白质转运是在翻译后进行的,在转运过程中没有蛋白质的合成。

有实验把豌豆的 RuBP 羧化酶小亚基 mRNA 加入麦胚无细胞蛋白质合成系统中,合成了相对分子质量为 $2.0\times10^4$ 的小亚基前体,离心处理含这种前体的麦胚提取物,除去核糖体,发现小亚基前体留在上清液中。如果把提纯的完整叶绿体加入这一上清液中,发现小亚基前体进入叶绿体基质内,降解为成熟小亚基(相对分子质量 $1.4\times10^4$),并与大亚基结合形成全酶。由此证实,小亚基前体是从核糖体上释放后开始转运,并在转运过程中发生蛋白质降解的。叶绿体蛋白质转运过程有如下特点:

① 活性蛋白水解酶位于叶绿体基质内,这是鉴别翻译后转运的指标之一。

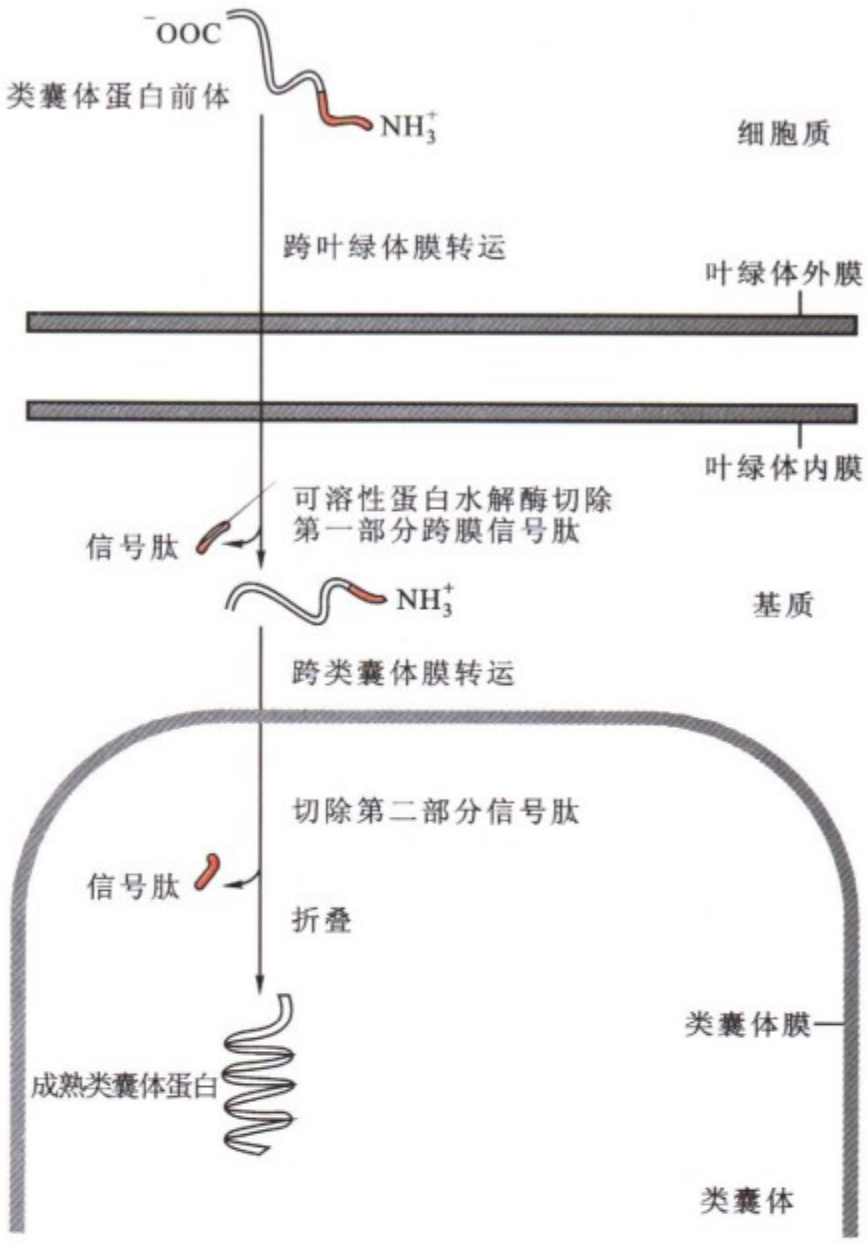



图 4-34 叶绿体蛋白质跨膜转运

从完整的叶绿体内提取的可溶性物质，能够把 RuBP 羧化酶小亚基前体降解或加工为成熟小亚基，离心后产生的叶绿体基质和破碎的叶绿体也具有这种功能，而类囊体和提纯的叶绿体膜都无此特性。在叶绿体蛋白质的翻译后转运机制中，活性蛋白酶是可溶性的，这一点也不同于分泌蛋白质的翻译-转运同步机制，因为后者活性蛋白酶位于转运膜上。因此，可根据蛋白水解酶的可溶性特征来区别这两种不同的转运机制。

② 叶绿体膜能够特异地与叶绿体蛋白的前体结合。从豌豆叶片中提取 poly A mRNA，置于麦胚提取物的上清液中合成蛋白质，并与分离的叶绿体膜共温育，发现 RuBP 羧化酶小亚基前体和聚光叶绿素 a、b 结合蛋白质的前体都能与叶绿体膜结合，而提取物中的其他蛋白质不与膜结合，说明叶绿体蛋白质前体与膜之间存在着特异性相互作用，或者说叶绿体膜上有识别叶绿体蛋白质的特异性受体。这种受体保证叶绿体蛋白质只能进入叶绿体内，而不是其他细胞器中。

③ 叶绿体蛋白质前体内可降解序列因植物和蛋白质种类不同而表现出明显的差异。如衣藻叶绿体中 RuBP 小亚基的可降解序列含有 44 个氨基酸残基，

在烟草叶绿体中相应的前导肽含有 57 个氨基酸残基。

### 4.5.3 核定位蛋白的转运机制

真核细胞的细胞核通过核孔与核外相通,因此,核孔是进行双向转运的分子通道。在细胞质中合成的蛋白质一般通过核孔进入细胞核。所有核糖体蛋白都首先在细胞质中被合成,转运到细胞核内,在核仁中被装配成 40S 和 60S 核糖体亚基,然后转运回到细胞质中行使作为蛋白质合成机器的功能。RNA、DNA 聚合酶、组蛋白、拓扑异构酶及大量转录、复制调控因子都必须从细胞质进入细胞核才能正常发挥功能。

在绝大部分多细胞真核生物中,每当细胞发生分裂时,核膜被破坏,等到细胞分裂完成后,核膜被重新建成,分散在细胞内的核蛋白必须被重新运入核内,因此,为了核蛋白的重复定位,这些蛋白质中的信号肽——被称为核定位序列(nuclear localization sequence,NLS)一般都不被切除。NLS 可以位于核蛋白的任何部位。

蛋白质向核内运输过程需要一系列循环于核内和细胞质的蛋白因子包括核运转因子(Importin)α、β 和一个低相对分子质量 GTP 酶(Ran)的参与。α 和 β 组成的异源二聚体是核定位蛋白的可溶性受体,与核定位序列相结合的是 α 亚基。由上述三个蛋白组成的复合物停靠在核孔处,依靠 Ran GTP 酶水解 GTP 提供的能量进入细胞核,α 和 β 亚基解离,核蛋白与 α 亚基解离,α 和 β 分别通过核孔复合体回到细胞质中,起始新一轮蛋白质转运(图 4-35)。

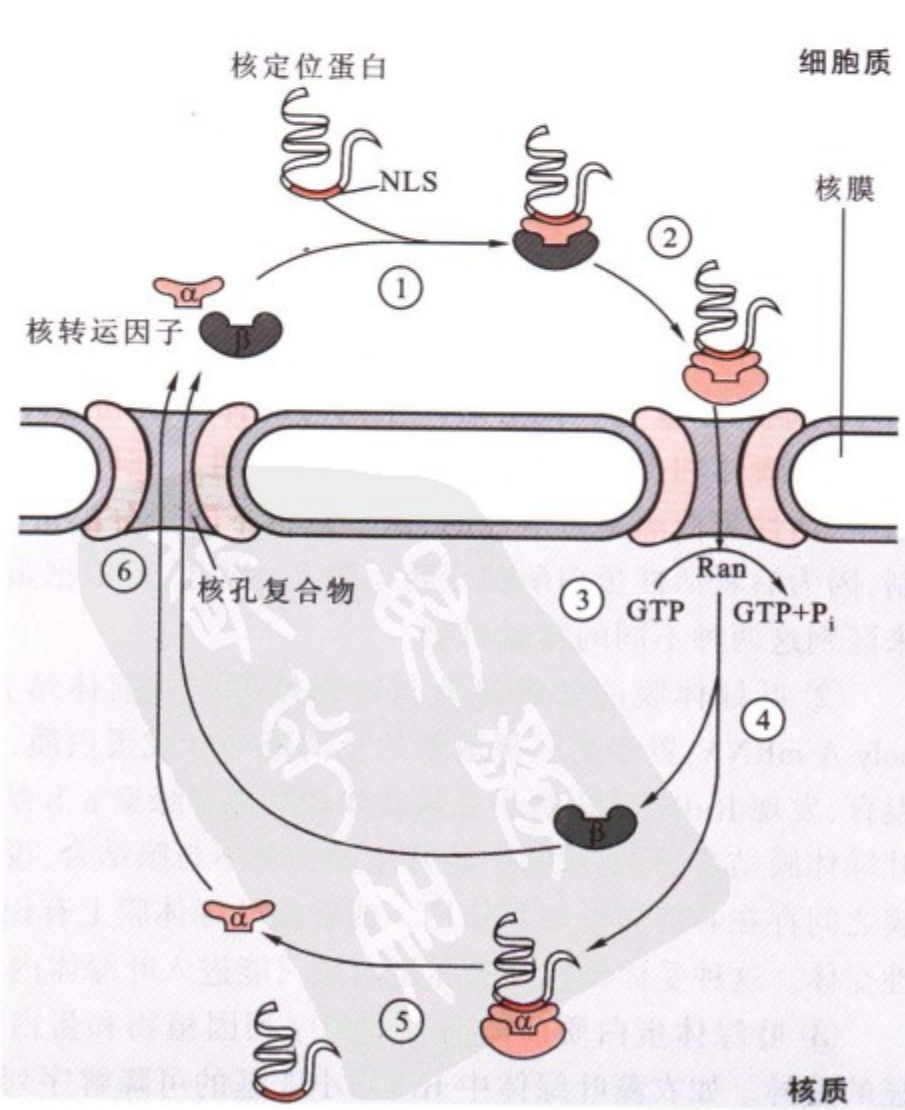



图 4-35 核定位蛋白跨细胞核膜转运过程示意图

对于原核细胞来说,同样存在蛋白质转运的问题。研究表明,细菌同样能通过定位于蛋白质 N 端的信号肽(图 4-36)将新合成的多肽转运到其内膜、外膜、双层膜之间或细胞外等不同部位。细菌中新翻译产生的蛋白质与胞质中的分子伴侣 SecB 相结合后就能被运送到细胞膜转运复合物 SecA-SecYEG 上,结合有新生肽的 SecA 在自身 ATP 酶活性作用下水解 ATP 并嵌入细胞膜之中,导致与 SecA 相结合的被转运蛋白 N 端约 20 个氨基酸通过膜转运复合物到达胞外。SecA 再与另一个 ATP 相结合,变构嵌入膜内的同时再次把所转运蛋白的第二部分运出胞外,如此反复,直到把所转运蛋白全部送到胞外(图 4-37)。

图 4-36 定位于不同亚细胞结构的细菌蛋白质信号肽序列比较

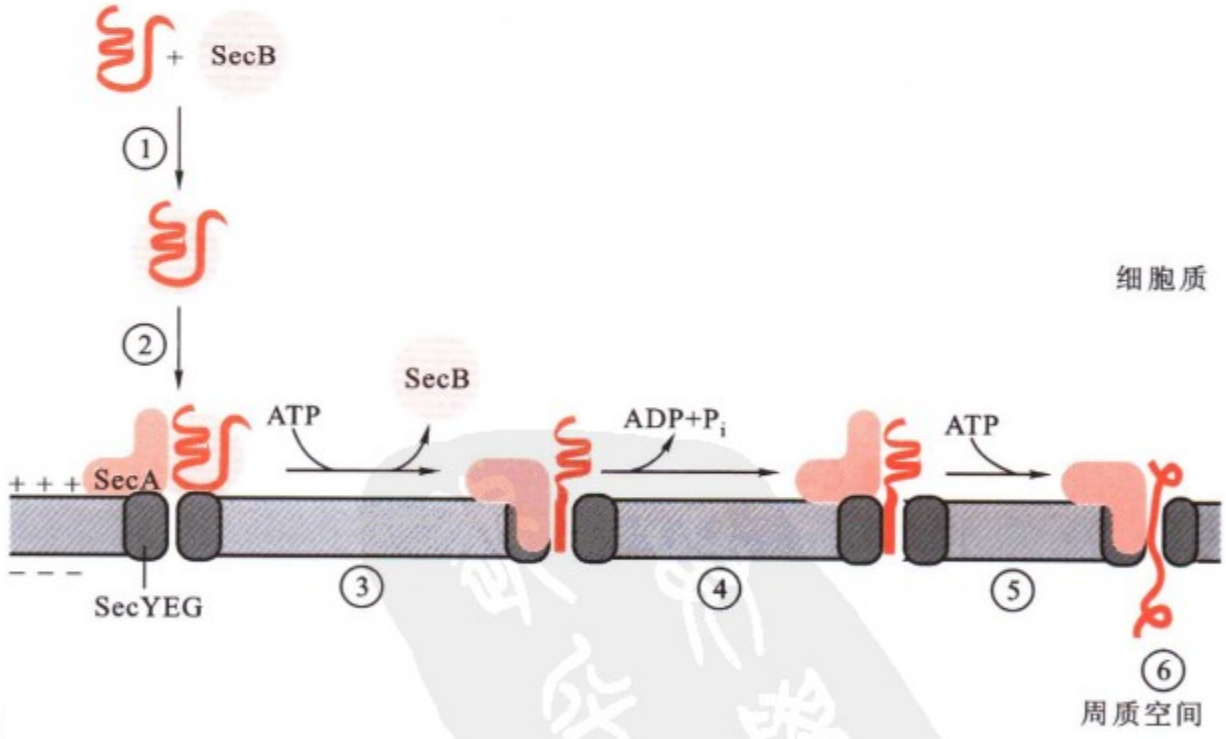



图 4-37 细菌中蛋白质的跨膜转运

## 4.5.4 蛋白质的降解

越来越多的证据表明,生物体内蛋白质的降解过程是一个有序的过程。在大肠杆菌中,蛋白质的降解是通过一个依赖于 ATP 的蛋白酶(称为 Lon)来实现的。当细胞中出现错误的蛋白质或半衰期很短的蛋白质时,该酶就被激活。Lon 蛋白酶每切除一个肽键要消耗两个分子的 ATP。

蛋白质的半衰期大约从 30 s 到许多天不等。虽然大部分真核蛋白的半衰期

为数小时至数天,像血红蛋白这样少数蛋白的半衰期与细胞周期同样长(成红细胞=110 d)。在真核生物中,蛋白质的降解依赖于泛蛋白(ubiquitin),该蛋白只有 76 个氨基酸残基,序列高度保守(从酵母和人细胞中提取的泛蛋白几乎完全相同!),生物细胞内即将被降解的蛋白首先在 ATP 的作用下与泛蛋白相连(图 4-38)。这个过程需有 E1、E2、E3 三个降解因子参与。一旦被降解蛋白与泛蛋白相连接,这个复合体就能被运送到相对分子质量高达 $1\times10^6$ 的蛋白降解体系中直到完全降解。

研究发现,成熟蛋白 N 端的第一个氨基酸(除已被切除的 N 端甲硫氨酸之外,但包括翻译后修饰产物)在蛋白的降解中有着举足轻重的影响(表 4-15)。当某个蛋白质的 N 端是甲硫氨酸、苷氨酸、丙氨酸、丝氨酸、苏氨酸和缬氨酸时,表现稳定。其 N 端为赖氨酸、精氨酸时,表现最不稳定,平均 2~3 min 就被降解了。泛蛋白调控的蛋白质降解具有重要的生理意义,它不仅能够清除错误蛋白

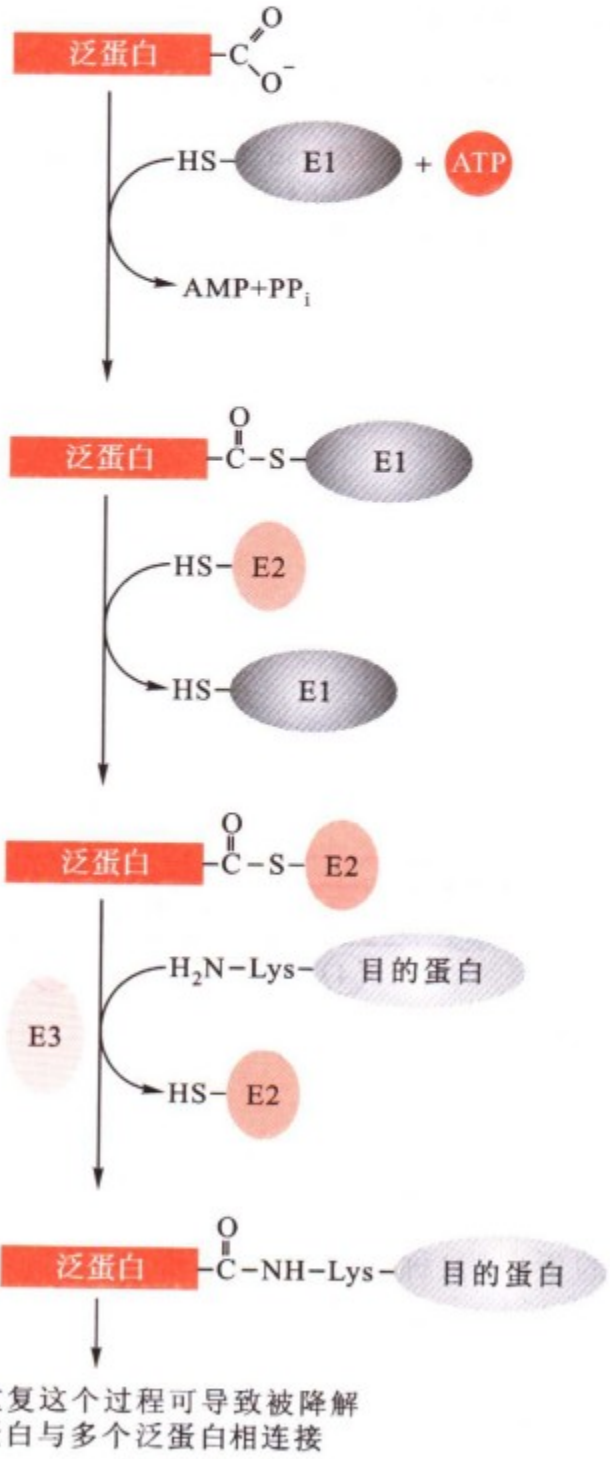



图 4-38 泛蛋白(泛素)在 E1、E2 和 E3 的作用下与被降解蛋白相连接的过程

质,对细胞生长周期、DNA 复制以及染色体结构都有重要的调控作用,而且对于理解细胞的许多生理过程和新药的开发也具有重要意义。

表 4-15 蛋白质的半衰期与 N 端氨基酸残基的关系

| N 端残基 | 半衰期 |
|---|---|
| 稳定型残基 | |
| 甲硫氨酸、苷氨酸、丙氨酸、丝氨酸、苏氨酸、缬氨酸 | >20 h |
| 不稳定型残基 | |
| 异亮氨酸、谷氨酰胺 | ~30 min |
| 酪氨酸、谷氨酸 | ~10 min |
| 脯氨酸 | ~7 min |
| 亮氨酸、苯丙氨酸、天冬酰胺、赖氨酸 | ~3 min |
| 精氨酸 | ~2 min |

## 思考题

1. 遗传密码有哪些特性?
2. 有几种终止密码子? 它们的序列和别名是什么?
3. 简述摆动学说。
4. tRNA 在组成和结构上有哪些特点?
5. 比较原核与真核的核糖体组成。
6. 什么是 SD 序列? 其功能是什么?
7. 核糖体有哪些活性中心?
8. 真核生物与原核生物在翻译的起始过程中有哪些区别?
9. 链霉素为什么能够抑制蛋白质的合成?
10. 什么是信号肽? 它在序列组成上有哪些特点? 有什么功能?
11. 简述叶绿体蛋白质的跨膜转运机制。
12. 蛋白质有哪些翻译后的加工修饰?
13. 什么是核定位序列? 其主要功能是什么?

## 主要参考文献


1. Algire M A, Lorsch J R. Where to begin? The mechanism of translation initiation codon selection in eukaryotes. Curr. Opin. Chem. Biol. 2006, 10: 480–486.
2. Carter A P, Clemons W M, Brodersen D E *et al.* Crystal structure of an initiation factor bound to the 30S ribosomal subunit. Science, 2001, 291: 498–501.
3. Ciechanover A. The ubiquitin proteolytic system: From a vague idea, through basic mechanisms, and onto human diseases and drug targeting. Neurology, 2006, 66: S7–19.

4. Crick F H C, Barnett L, Brenner S *et al*. General nature of the genetic code for proteins. Nature, 1961, 192: 1227–1232.
5. Crick F H C. Codon–anticodon pairing: the wobble hypothesis. J. Mol. Biol. 1966, 19: 548–555.
6. Cusack S. Aminoacyl-tRNA synthetases. Curr. Opin. Struct. Biol. 1997, 7: 881–889.
7. Freistroffer D V, Kwiatkowski M, Buckingham R H, Ehrenberg M. The accuracy of codon recognition by polypeptide release factors. Proc. Natl. Acad. Sci. U.S.A. 2000, 97: 2046–2051.
8. Hopper A K, Phizicky E M. tRNA transfers to the limelight. Genes Dev. 2003, 17: 162–180.
9. Lafontaine D L, Tollervey D. The function and synthesis of ribosomes. Nat. Rev. Mol. Cell Biol, 2001, 2: 514–520.
10. Mitra K, Frank J, Driessen A. Co- and post-translational translocation through the pprotein-conducting channel: analogous mechanisms at work? Nat. Struct. Mol. Biol. 2006, 13: 957–964.
11. Nirenberg M W, Matthaei H J. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc. Natl. Acad. Sci. U.S.A. 1961, 47: 1588–1602.
12. Ramakrishnan V. Ribosome structure and the mechanism of translation. Cell, 2002, 108: 557–572.

# 5 分子生物学研究法(上)
## ——DNA、RNA 及蛋白质操作技术


- 5.1 重组 DNA 技术回顾
- 5.2 DNA 基本操作技术
- 5.3 RNA 基本操作技术
- 5.4 SNP 的理论与应用
- 5.5 基因克隆技术
- 5.6 蛋白质组与蛋白质组学技术


分子生物学研究之所以从 20 世纪中叶开始得到高速发展，最主要的原因之一就是现代分子生物学研究方法，特别是基因操作和基因工程技术的进步。基因操作主要包括 DNA 分子的切割与连接、核酸分子杂交、凝胶电泳、细胞转化、核酸序列分析以及基因的人工合成、定点突变和 PCR 扩增等，是分子生物学研究的核心技术。基因工程是指在体外将核酸分子插入病毒、质粒或其他载体分子，构成遗传物质的新组合，使之进入新的宿主细胞内并获得持续稳定增殖能力和表达。因此，基因工程技术其实际上是核酸操作技术的一部分，只不过我们在这里强调了外源核酸分子在另一种不同的宿主细胞中的繁衍与性状表达。事实上，这种跨越天然物种屏障、把来自任何生物的基因置于毫无亲缘关系的新的宿主生物细胞之中的能力，是基因工程技术区别于其他技术的根本特征。本章将在回顾重组 DNA 技术发展史的基础上，讨论 DNA 操作技术、基因克隆、表达分析技术及蛋白质组学、单核苷酸多态性分析等现代生物学领域里最广泛应用的实验技术和方法。

## 5.1 重组 DNA 技术回顾

近半个多世纪来，分子生物学研究取得了前所未有的进步，概括地说，主要有三大成就：第一，在 20 世纪 40 年代确定了遗传信息的携带者，即基因的分子载体是 DNA 而不是蛋白质，解决了遗传的物质基础问题；第二，50 年代提出了 DNA 分子的双螺旋结构模型和半保留复制机制，解决了基因的自我复制和世代交替问题；第三，50 年代末至 60 年代，相继提出了“中心法则”和操纵子

学说,成功地破译了遗传密码,阐明了遗传信息的流动与表达机制。但是,由于缺乏有效的分离和富集单一 DNA 分子的技术,科学家无法对这类物质进行直接的生物化学分析。事实上,DNA 分子体外切割与连接技术及核苷酸序列分析技术的进步直接推动了重组 DNA 技术的产生与发展。

因为重组 DNA 的核心是用限制性内切核酸酶 (restriction endonuclease, RE)和 DNA 连接酶对 DNA 分子进行体外切割与连接,所以,科学家认为,这些工具酶的发现和应用是现代生物工程技术史上最重要的事件(表 5-1)。

表 5-1 重组 DNA 技术史上的主要事件

| 年份 | 事 件 |
|---|---|
| 1869 | F. Miescher 首次从莱茵河鲑鱼精子中分离 DNA |
| 1957 | A. Kornberg 从大肠杆菌中发现了 DNA 聚合酶 I |
| 1959—1960 | Uchoa 发现 RNA 聚合酶和信使 RNA,并证明 mRNA 决定了蛋白质分子中的氨基酸序列 |
| 1961 | Nirenberg 破译了第一个遗传密码;Jacob 和 Monod 提出了调节基因表达的操纵子模型 |
| 1964 | Yanofsky 和 Brenner 等人证明,多肽链上的氨基酸序列与该基因中的核苷酸序列存在着共线性关系 |
| 1965 | Holley 完成了酵母丙氨酸 tRNA 的全序列测定;科学家证明细菌的抗药性通常由“质粒”DNA 所决定 |
| 1966 | Nirenberg, Uchoa, Khorana, Crick 等人破译了全部遗传密码 |
| 1967 | 第一次发现 DNA 连接酶 |
| 1970 | Smith, Wilcox 和 Kelley 分离了第一种限制性内切核酸酶, Temin 和 Baltimore 从 RNA 肿瘤病毒中发现反转录酶 |
| 1972—1973 | Boyer, Berg 等人发展了 DNA 重组技术,于 1972 年获得第一个重组 DNA 分子,1973 年完成第一例细菌基因克隆 |
| 1975—1977 | Sanger 与 Maxam 和 Gilbert 等人发明了 DNA 序列测定技术,1977 年完成了全长 5 387 bp 的噬菌体 φ174 基因组测定 |
| 1978 | 首次在大肠杆菌中生产由人工合成基因表达的人脑激素和人胰岛素 |
| 1980 | 美国联邦最高法院裁定微生物基因工程可以被专利化 |
| 1981 | Palmiter 和 Brinster 获得转基因小鼠, Spradling 和 Rubin 得到转基因果蝇 |
| 1982 | 美、英批准使用第一例基因工程药物——胰岛素, Sanger 等人完成了 λ 噬菌体 48 502 bp 全序列测定 |
| 1983 | 获得第一例转基因植物 |
| 1984 | 斯坦福大学获得关于重组 DNA 的专利 |
| 1986 | GMO 首次在环境中释放 |

续表

| 年份 | 事　件 |
|---|---|
| 1988 | Watson 出任“人类基因组计划”首席科学家 |
| 1989 | DuPont 公司获得转肿瘤基因小鼠——“Oncomouse” |
| 1992 | 欧共体 35 个实验室联合完成酵母第三染色体全序列测定(315 kb) |
| 1994 | 第一批基因工程西红柿在美国上市 |
| 1996 | 完成了酵母基因组($1.25 \times 10^7$ bp)全序列测定 |
| 1997 | 英国爱丁堡罗斯林研究所获得克隆羊 |
| 2000 | 完成第一个高等植物拟南芥的全序列测定($1.2 \times 10^8$ bp) |
| 2001 | 完成第一个人类基因组全序列测定($2.7 \times 10^9$ bp) |
| 2004 | 中国科学家完成家蚕基因组全序列测定 |
| 2005 | 中、美、日科学家联合完成水稻基因组序列测定 |
| 2005 | 美、德、日、意、以科学家完成黑猩猩基因组全序列测定 |
| 2003 | 美国科学家完成狗基因组全序列测定 |

限制性内切核酸酶能够识别 DNA 上的特定碱基序列并从这个位点切开 DNA 分子。1972 年,Boyer 实验室发现有一种内切核酸酶 *Eco*R Ⅰ能特异性识别 GAATTC 序列,将双链 DNA 分子在这个位点切开并产生具有黏性末端的小片段。他们还发现,能够把 *Eco*R Ⅰ酶切产生的任何不同来源的 DNA 片段通过黏性末端之间的碱基互补作用而彼此“黏合”起来。此后,大量类似于 *Eco*R Ⅰ但具有独特识别序列的内切核酸酶被陆续发现(图 5-1)。截至目前,

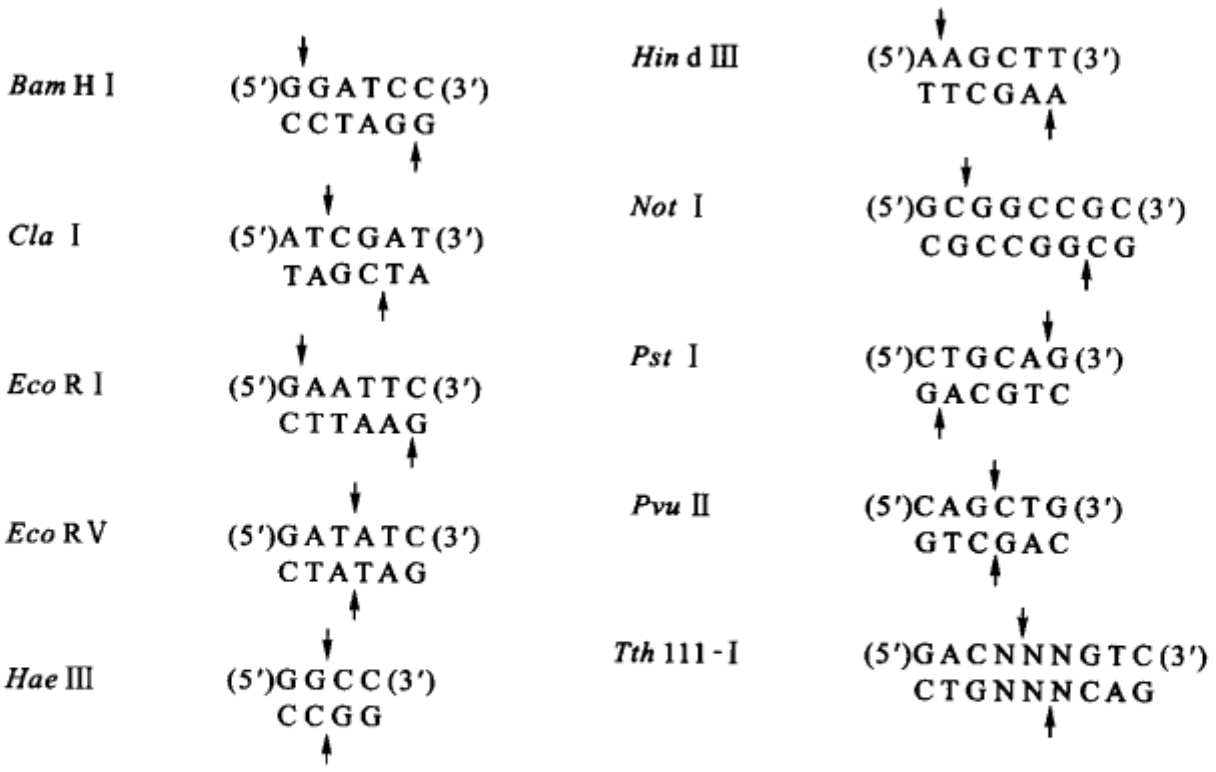



图 5-1　几种主要 DNA 内切酶所识别的序列及其酶切位点

科学家已经几乎能随心所欲地把任何DNA分子切割成一系列不连续的片段,再利用凝胶电泳技术将这些片段按照相对分子质量大小逐一分开,供下一步研究。

1967年,世界上有5个实验室几乎同时报道了通过合成相邻核苷酸之间的磷酸二酯键,从而修复缺口或催化黏合完全分离的两个DNA片段的DNA连接酶(图5-2)。当然,仅仅能在体外利用限制性内切核酸酶和DNA连接酶进行DNA的切割和重组,还远远不能满足基因工程的要求,因为大多数DNA片段不具备自我复制的能力,只有将它们连接到具备自主复制能力的DNA分子上,才能在宿主细胞中进行繁殖。这种具备自主复制能力的DNA分子就是所谓分子克隆的载体(vector),病毒、噬菌体和质粒等小相对分子质量复制子都可以作为基因导入的载体。

获得了用外源DNA片段和载体分子重组而成的杂种DNA分子后,还必须通过一个被称为细菌转化的过程将其重新导入到宿主细胞中,才能保证重组DNA分子的增殖(图5-3)。1970年,Mandel和Higa发现,大肠杆菌细胞经适量氯化钙处理后,能有效地吸收λ噬菌体DNA。1972年,Cohen等人又报道,经氯化钙处理的大肠杆菌细胞同样能够摄取质粒DNA。从此,大肠杆菌就成了分子克隆中最常用的转化受体。同年,美国斯坦福大学的Berg等人在体外用限制性内切核酸酶*Eco*R Ⅰ分别消化猿猴病毒SV40和λ噬菌体DNA,再用T4 DNA连接酶将两种不同的消化片段连接起来,首次获得了同时含有SV40和λ噬菌体DNA的重组DNA。1973年,Cohen实验室也成功地进行了体外DNA重组,他们将分别经限制性内切核酸酶处理的编码卡那霉素抗性基因的大肠杆菌质粒R6-5 DNA与编码四环素抗性基因的另一种大肠杆菌质粒pSC101 DNA混合后加入T4 DNA连接酶,得到的重组杂种DNA分子表现出既抗卡那霉素又抗四环素的双重抗性特征。

R6-5和pSC101都是来自大肠杆菌的质粒DNA,由它们形成的重组DNA分子可以在原来的宿主细胞中增殖,这似乎还比较容易理解。那么,不同物种的外源DNA片段是否也可以在大肠杆菌细胞中增殖呢?为了回答这个问题,Cohen和Boyer等人合作,把非洲爪蟾核糖体蛋白基因片段与pSC101质粒DNA片段重组后导入大肠杆菌,证明动物基因也能进入大肠杆菌细胞并转录出相应的mRNA分子。上述实验表明:

(1)像pSC101这样的质粒DNA分子可以作为基因克隆的载体,从而把外源DNA导入宿主细胞。

(2)像非洲爪蟾这样的高等生物的基因也可以被成功地转移到原核细胞中并实现其功能表达。

(3)质粒DNA-大肠杆菌细胞作为一种成功的基因克隆体系,有可能在重组DNA或基因工程研究中发挥重要作用。

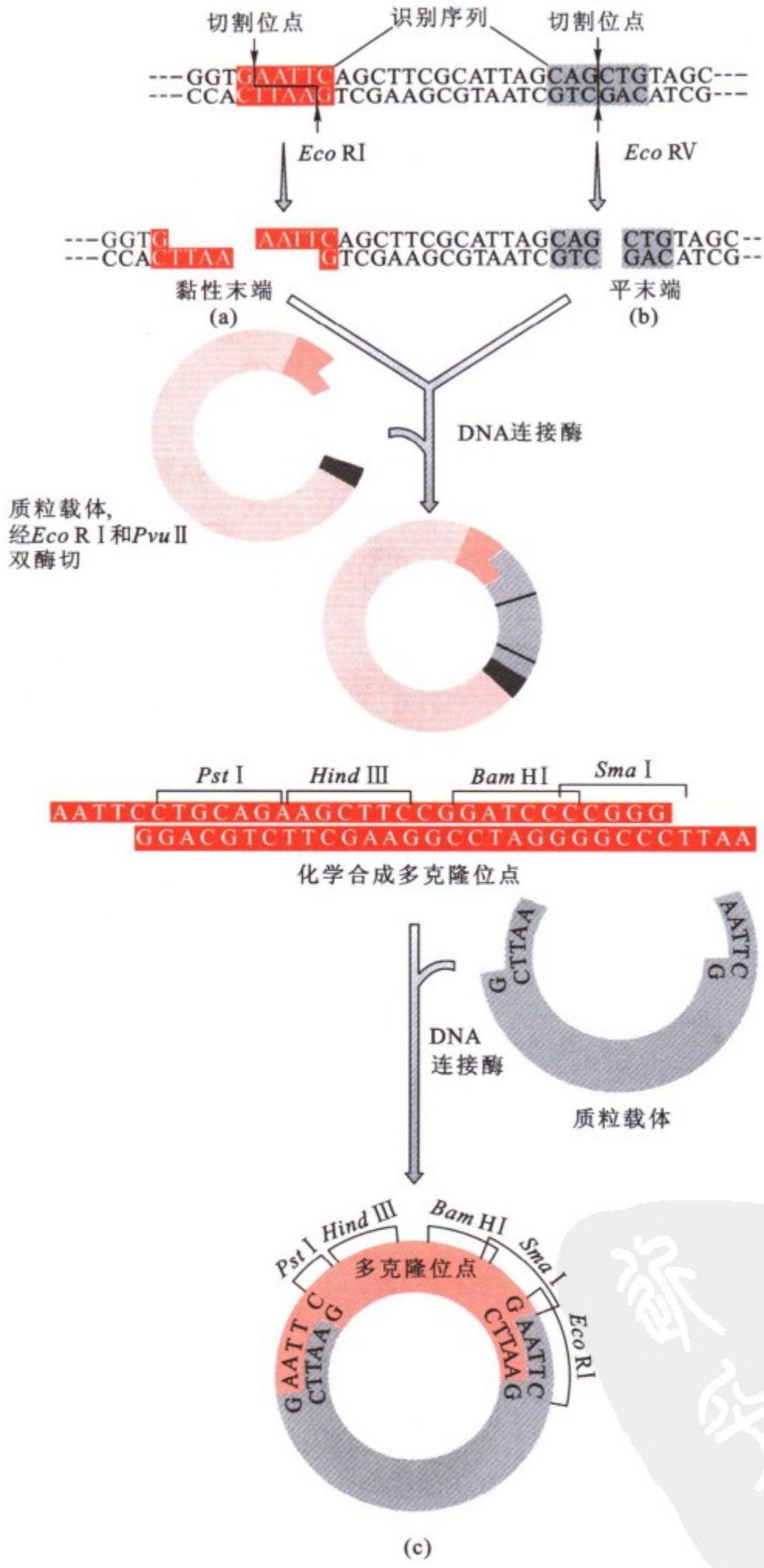
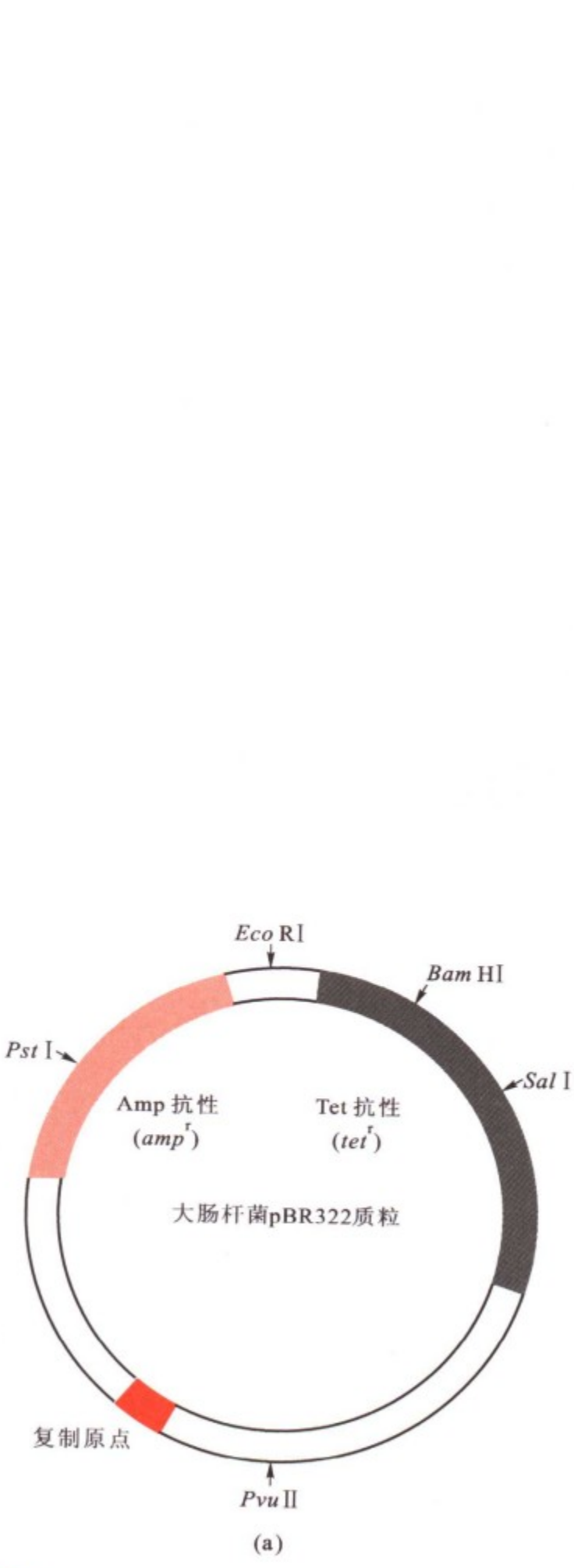



**图 5–2 DNA 连接酶能把不同的 DNA 片段连接成一个整体** (a)DNA 的黏性末端;(b)DNA 的平末端;(c) 化学合成的具有 *Eco* R Ⅰ 黏性末端的 DNA 片段。



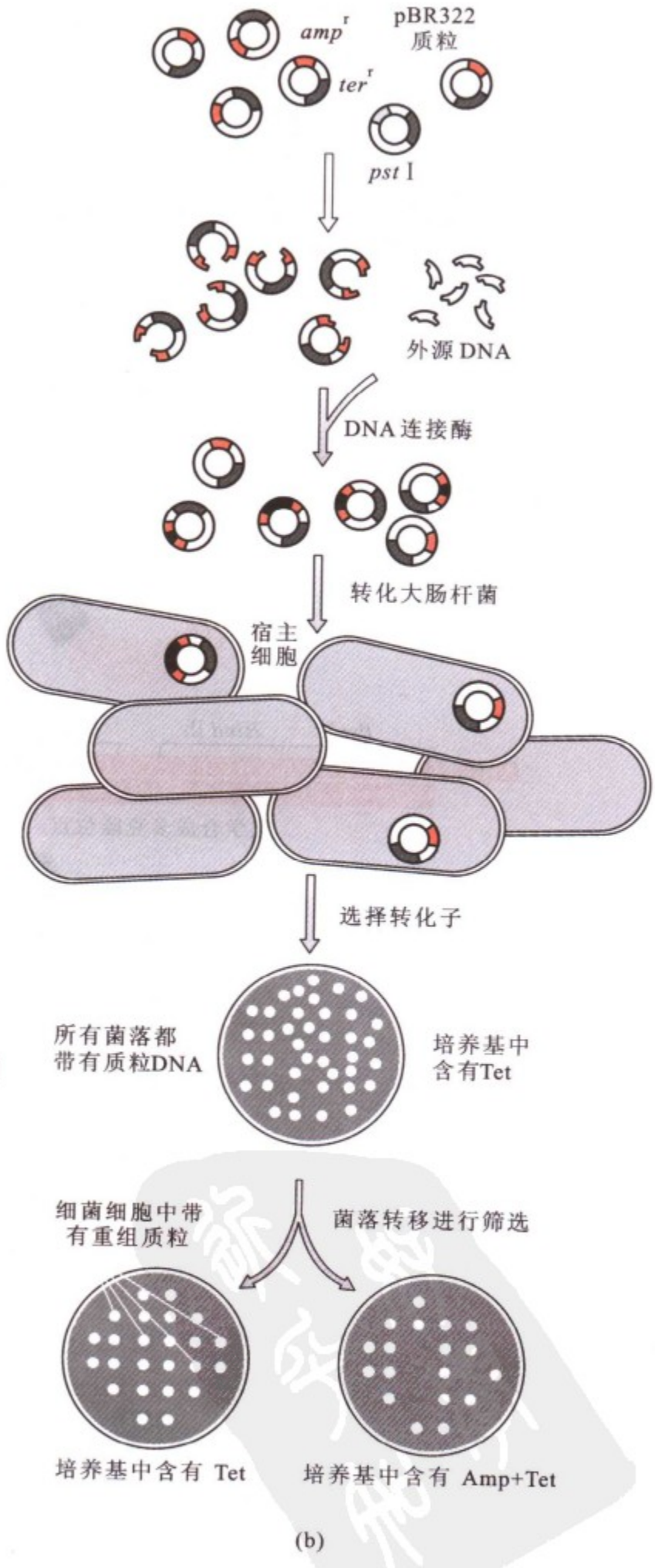




图 5-3 重组 DNA 操作过程示意图

# 5.2 DNA 基本操作技术

## 5.2.1 核酸凝胶电泳技术

自从琼脂糖(agarose)和聚丙烯酰胺(polyacrylamide)凝胶被引入核酸研究以来,按相对分子质量大小分离 DNA 的凝胶电泳技术,已经发展成为一种分析鉴定重组 DNA 分子及蛋白质与核酸相互作用的重要实验手段,也是现在通用的许多分子生物学研究方法,如 DNA 分型、DNA 核苷酸序列分析、限制性内切酶片段分析以及限制酶切作图等的技术基础,受到科学界的高度重视。

实验中,若将一种分子置于电场中,它就会以一定的速度向适当的电极移动,我们把这种电泳分子在电场作用下的迁移速度,叫做电泳的迁移率,它与电场强度和电泳分子本身所携带的净电荷数成正比。就是说,电场强度越大,电泳分子所携带的净电荷数量越多,其迁移的速度越快,反之则较慢。由于在电泳中往往使用无反应活性的稳定的支持介质,如琼脂糖凝胶和聚丙烯酰胺凝胶等,故电泳的迁移率与分子的摩擦系数成反比。已知摩擦系数是分子大小、极性及介质黏度的函数。因此,根据分子大小的不同、构型或形状的差异以及所带的净电荷的多寡,便可以通过电泳将蛋白质或核酸分子混合物中的各种成分彼此分离开来。

在生理条件下,核酸分子之糖-磷酸骨架中的磷酸基团,是呈离子化状态的,所以,DNA 和 RNA 多核苷酸链又被称为多聚阴离子(polyanions)。把这些核酸分子放置在电场当中,它们就会向正电极的方向迁移。由于糖-磷酸骨架在结构上的重复性质,相同数量的双链 DNA 几乎具有等量的净电荷,因此它们能以同样的速度向正电极方向迁移。在一定的电场强度下,DNA 分子的这种迁移速度,亦即电泳的迁移率,取决于核酸分子本身的大小和构型。这就是应用凝胶电泳技术分离 DNA 片段的基本原理。

琼脂糖是一种线性多糖聚合物,从红色海藻产物琼脂中提取而来。将琼脂糖粉末加热到熔点后冷却凝固便会形成良好的电泳介质,其密度由琼脂糖的浓度决定。经过化学修饰的低熔点琼脂糖,在结构上比较脆弱,因此,在较低的温度下便会熔化,常用于 DNA 片段的制备电泳。

凝胶的分辨能力同凝胶的类型和浓度有关(表 5-2)。琼脂糖凝胶分辨 DNA

表 5-2 琼脂糖及聚丙烯酰胺凝胶分辨 DNA 片段的能力

| 凝胶类型及浓度 | 分离 DNA 片段的大小范围/bp |
|---|---|
| 0.3% 琼脂糖 | 50 000 ~ 1 000 |
| 0.7% 琼脂糖 | 20 000 ~ 1 000 |
| 1.4% 琼脂糖 | 6 000 ~ 300 |
| 4.0% 聚丙烯酰胺 | 1 000 ~ 100 |
| 10.0% 聚丙烯酰胺 | 500 ~ 25 |
| 20.0% 聚丙烯酰胺 | 50 ~ 1 |

片段的范围为0.2~50 kb;而要分辨较小相对分子质量的DNA片段,则要用聚丙烯酰胺凝胶,其分辨范围为1~1 000 bp。凝胶浓度的高低影响凝胶介质孔隙的大小,浓度越高,孔隙越小,其分辨能力就越强。反之,浓度降低,孔隙就增大,其分辨能力也就随之减弱。例如,20%的聚丙烯酰胺凝胶可分离1~6 bp的DNA小片段,而若要分离1 000 bp的DNA片段,就要用3%的聚丙烯酰胺凝胶。再如,2%的琼脂糖凝胶可分离300 bp的双链DNA分子,而分离大片段DNA就要用低浓度(0.3%~1.0%)的琼脂糖凝胶。

在凝胶电泳中,加入适量溴乙啶(ethidium bromide,EB)染料对核酸分子进行染色,然后将电泳标本放置在紫外线下观察,可十分灵敏而快捷地检测出凝胶介质中DNA的谱带部位,即使每条DNA带中仅含有0.05 μg的微量DNA,也可以被清晰地检测出来。溴乙啶是一种具扁平分子的核酸染料,能插入到DNA或RNA分子的相邻碱基之间,并在300 nm波长的紫外线照射下发出荧光。把含有DNA分子的凝胶浸泡在溴乙锭的溶液中,或是将溴乙啶直接加到DNA的凝胶介质中,此种染料便会在一切可能的部位与DNA分子结合,然而却不能与琼脂糖凝胶或聚丙烯酰胺凝胶相结合,因此,只有DNA分子能吸收溴乙啶并发出荧光。在适当的染色条件下,荧光强度与DNA片段的大小(或数量)成正比。

实验表明,用普通琼脂糖凝胶电泳,很难分离大于50kb的DNA分子,为了进行超大片段DNA研究,科学家发明了脉冲电场凝胶电泳(pulsed-field gel electrophoresis,PFGE)技术,用于分离超大相对分子质量(有时甚至是整条染色体)DNA。在脉冲电场中,DNA分子的迁移方向随着电场方向的周期性变化而不断改变。在标准的PFGE中,第一个脉冲的电场方向与核酸移动方向成45°夹角,而第二个脉冲的电场方向与核酸移动方向在另一侧成45°夹角(图5-4)。由于琼脂糖凝胶的电场方向、电流大小及作用时间都在交替变换,使得DNA分子能够随时调整其游动方向,以适应凝胶孔隙的无规则变化。与相对分子质量较小的DNA相比,相对分子质量较大的DNA需要更多的脉冲次数来更换其构型和方位,使其按新的方向游动。应用脉冲电场凝胶电泳技术,可成功地分离到相对分子质量高达$10^7$ bp的DNA大分子。

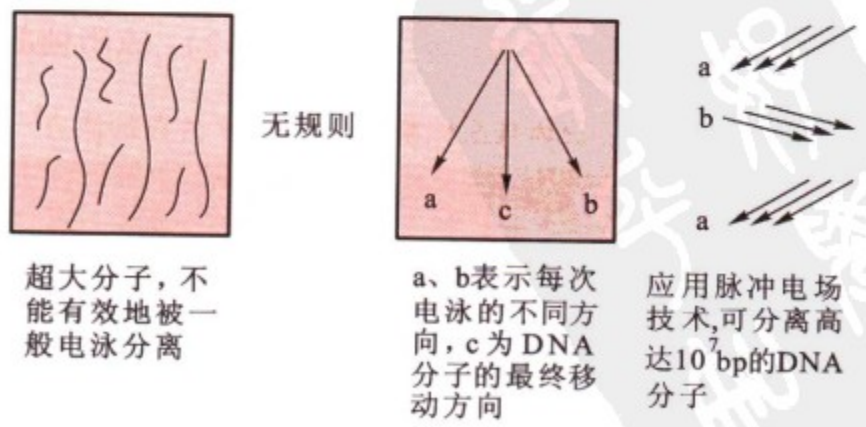



图5-4 DNA脉冲电场凝胶电泳示意图

## 5.2.2 细菌转化与目标 DNA 分子的增殖

细菌转化(transformation),是指一种细菌菌株由于捕获了来自供体菌株的 DNA 而导致性状特征发生遗传改变的过程。提供转化 DNA 的菌株叫做供体菌株,接受转化 DNA 的细菌菌株则被称为受体菌株。

在基因克隆的过程中,我们通常要通过细菌转化将体外构建好的杂种 DNA 分子导入宿主细胞中。进入宿主细胞的外源 DNA 通过自身载体上的复制起始点进行复制增殖,使其能够在宿主细胞中保持下来,并且还能够很容易地以完整的形式从细胞中被分离纯化出来。

绝大多数细菌,包括大肠杆菌 (*E.coli*),正常条件下仅能获取极少量的 DNA。为了高效转化这些细菌,必须对受体细菌进行一些物理或化学处理,以增加它们获取 DNA 的能力。经历了这种处理的细胞被称作感受态细胞(competent cells)。细菌转化常用的方法有 $CaCl_2$ 法和电击法。

$CaCl_2$ 法是制备大肠杆菌感受态最常用的化学方法。将快速生长的大肠杆菌置于经低温(0℃)预处理的低渗 $CaCl_2$ 溶液中,便会造成细胞膨胀,细胞膜通透性发生变化,极易与外源 DNA 相黏附并在细胞表面形成复合物。此时,将该体系转移到 42℃下做短暂的热刺激,外源 DNA 就可能被细胞吸收。进入细胞的外源 DNA 分子通过复制、表达,实现遗传信息的转移,使受体细胞出现新的遗传性状。将经过转化后的细胞在选择性培养基上培养,筛选出带有外源 DNA 分子的阳性克隆。这一方法可以用于批量制备感受态细胞,其转化效率可达到 $5\times10^6$~$2\times10^7$ 个转化子/μg 超螺旋质粒 DNA(图 5-5)。

电击法是细菌转化的另一种高效方法。锐利的电脉冲可以在细胞膜上造成小凹陷,并由此形成纳米级疏水孔洞。随着跨膜电压增加,一些较大的疏水性孔洞会转变为亲水性孔洞。在孔洞开放的时候,介质中的 DNA 很容易通过孔洞进入细胞质。将生长至对数中期的 *E.coli* 菌液冷却至 4℃后离心,用相同温度的纯水洗涤,用 10%的甘油悬浮后将高密度菌液(~$2\times10^{10}$/mL)置于特制的电极杯中进行电击。获得最大转化效率时场强一般为 12.5~15 kV/cm,时间跨度一般为 4.5~5.5 ms。电击转化与温度有关,最好在 0~4℃进行。由于转化载体上一般带有 *Lac*Z 基因,实验室常用带有不同抗生素(常用的抗生素包括氨苄西林、卡那霉素和四环素等)的选择性培养基结合 α-互补蓝白斑筛选法鉴定转化细胞。感受态细胞转化频率的计算方法为:

转化体总数=菌落数×(转化反应原液总体积/涂板菌液体积)

插入频率=白色菌落数/(白色菌落数+蓝色菌落数)

转化频率(每 μg DNA 转化菌落数)=转化体总数/加入质粒 DNA 总量(μg)

利用噬菌体颗粒能够有效地将 DNA 注入宿主细胞中这一特点,科学家又发明了重组体 DNA 分子的体外包装法(图 5-6)。经包装的基因工程噬菌体颗粒能够借助细菌表面的噬菌体接受器位点将 DNA 注入受体细菌,并在这些细胞中得到繁殖和表达。

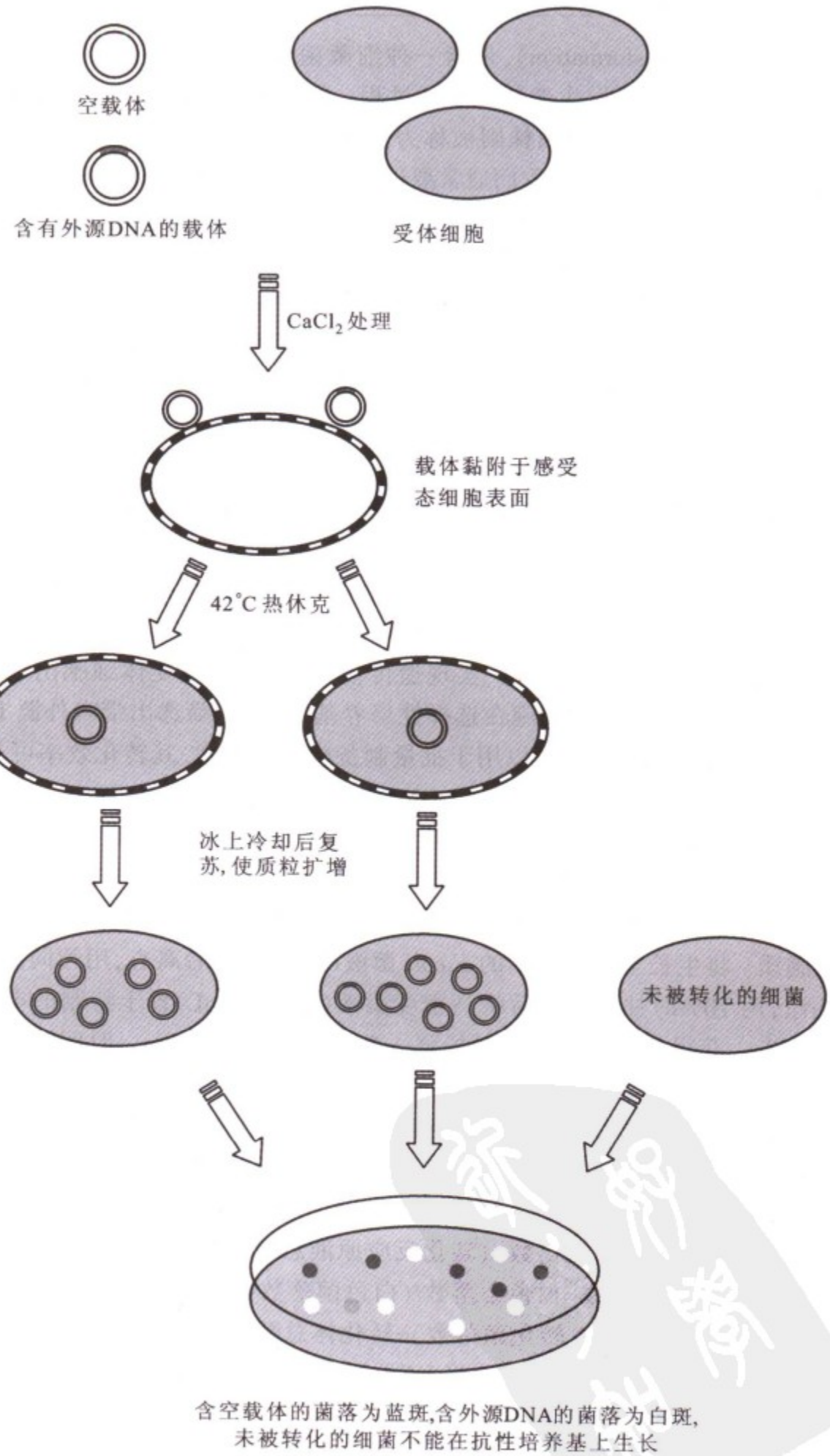



图 5-5 细菌转化及蓝白斑筛选

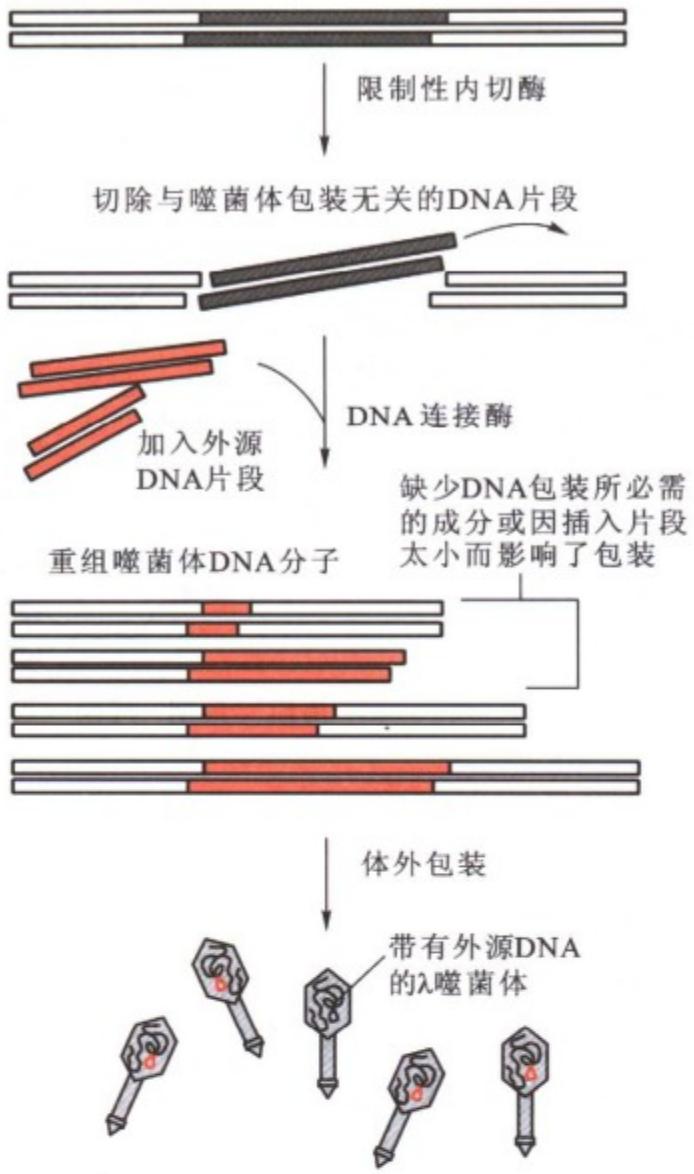



图 5-6 λ噬菌体作为克隆载体构建基因文库的流程图

### 5.2.3 聚合酶链反应技术

聚合酶链反应(PCR)技术可能是体外快速扩增特定基因或 DNA 序列最常用的方法。20 世纪 80 年代,凯利·穆利斯(Kary Mullis)发明了该技术。当时,他供职于美国的西特斯公司——世界上第一间靠重组 DNA 技术起家的公司。PCR 反应的起始材料,模板 DNA,可以是基因组 DNA 上的某个基因或基因片段(genomic PCR),也可以是 mRNA 反转录产生的 cDNA 链(RT-PCT)。PCR 技术的原理并不复杂。首先将双链 DNA 分子在临近沸点的温度下加热分离成两条单链 DNA 分子,DNA 聚合酶以单链 DNA 为模板并利用反应混合物中的四种脱氧核苷三磷酸合成新生的 DNA 互补链(图 5-7)。DNA 聚合酶是一种天然产生的能催化 DNA(包括 RNA)的合成和修复的生物大分子。所有生物体基因组的准确复制都依赖于这个酶的活性。PCR 反应中使用的 DNA 聚合酶不同于一般的聚合酶,它具有很好的耐高温性。DNA 聚合酶开始 DNA 合成时都需要有一小段双链 DNA 来启动("引导")新链的合成,所以,新合成的 DNA 链的起点,事实上是由加入到反应混合物中的一对寡核苷酸引物在模板 DNA 链两端的退火位点

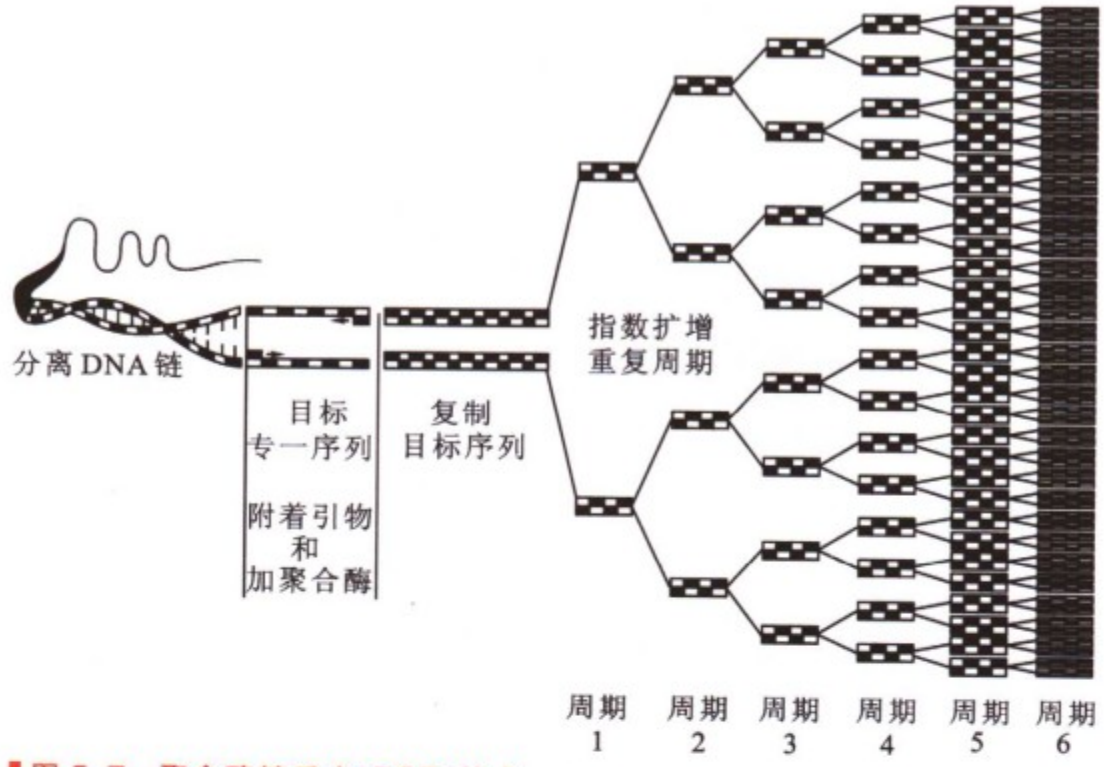



图 5-7 聚合酶链反应(PCR)技术

决定的。

PCR反应时，只要在试管内加入模板DNA、PCR引物、四种核苷酸及适当浓度的$Mg^{2+}$，DNA聚合酶就能在数小时内将目标序列扩增100万倍以上。双链模板DNA分子首先在高温下解开成长的单链，短链引物分子立即与该模板DNA两端的特定序列相结合，产生双链区。DNA聚合酶从引物处开始复制其互补链，迅速产生与目标序列完全相同的复制品。在后续反应中，无论是起始模板DNA还是经复制的杂合DNA双链，都会在高温下解开成为单链，体系中的引物分子再次与其互补序列相结合，聚合酶也再度复制模板DNA。由于在PCR反应中所选用的一对引物，是按照与扩增区段两端序列彼此互补的原则设计的，因此，每一条新生链的合成都是从引物的退火结合位点开始并朝相反方向延伸的，每一条新合成的DNA链上都具有新的引物结合位点。整个PCR反应的全过程，即DNA解链(变性)、引物与模板DNA相结合(退火)、DNA合成(链的延伸)三步，可以被不断重复。经多次循环之后，反应混合物中所含有的双链DNA分子数，即两条引物结合位点之间的DNA区段的拷贝数，理论上的最高值应是$2^n$(图5-8)，能满足进一步遗传分析的需要。

| 周期 | 拷贝 |
|---|---|
| 1 | 2 |
| 2 | 4 |
| 3 | 8 |
| 4 | 16 |
| 5 | 32 |
| 6 | 64 |
| 7 | 128 |
| 8 | 256 |
| 9 | 512 |
| 10 | 1 024 |
| 11 | 2 048 |
| 12 | 4 096 |
| 13 | 8 192 |
| 14 | 16 384 |
| 15 | 32 788 |
| 16 | 65 536 |
| 17 | 131 072 |
| 18 | 262 144 |
| 19 | 524 288 |
| 20 | 1 048 576 |
| 21 | 2 097 152 |
| 22 | 4 194 304 |
| 23 | 8 388 608 |
| 24 | 16 777 216 |
| 25 | 33 554 432 |
| 26 | 67 108 864 |
| 27 | 134 217 728 |
| 28 | 268 435 456 |
| 29 | 536 870 912 |
| 30 | 1 073 741 824 |

图 5-8 PCR指数扩增时循环次数与DNA产物数量的比较

## 5.2.4 实时定量 PCR

由于 PCR 技术具有极高的敏感性，扩增产物总量的变异系数常常达到 10%~30%。因此，人们普遍认为应用简单方法对 PCR 扩增的产物进行最终定量是不可靠的。随着技术的进步，20 世纪 90 年代末期出现了实时定量 PCR(real time quantitative PCR)技术，利用带荧光检测的 PCR 仪对整个 PCR 过程中扩增 DNA 的累积速率绘制动态变化图，从而消除了在测定终端产物丰度时有较大变异系数的问题。

实时定量 PCR 反应在带透明盖的塑料小管中进行，激发光可以直接透过管盖，使其中的荧光探针被激发。荧光探针事先混合在 PCR 反应液中，只有与 DNA 结合后，才能够被激发出荧光。随着新合成目的 DNA 片段的增加，结合到 DNA 上的荧光探针，即被激发产生的荧光相应增加。最简单的 DNA 结合的荧光探针是非序列特异性的，例如，荧光染料 SYBR Green Ⅰ，激发光波长 520 nm，这种荧光染料只能与双链 DNA 结合，并且只有在与 DNA 结合时，才能被激发产生荧光(图 5-9)。

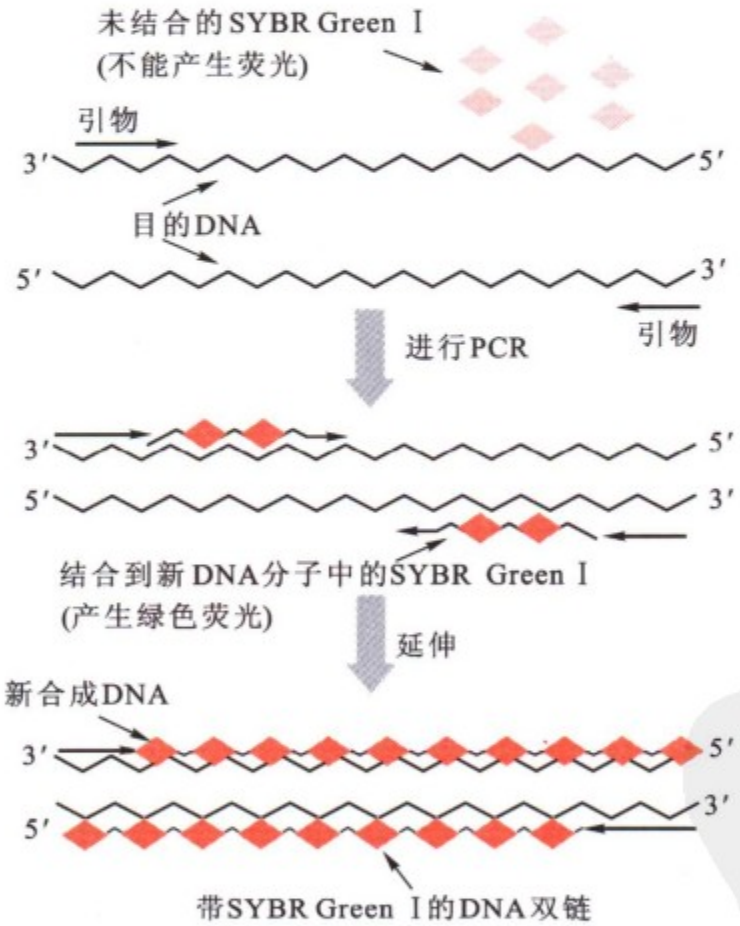



**图 5-9 SYBR Green Ⅰ做探针的实时定量 PCR 实验过程图示** 加入到 PCR 反应体系中的荧光染料 SYBR Green Ⅰ仅能与双链 DNA 结合，被激发出绿色荧光，其荧光强度反映 PCR 产物的产量。

利用 SYBR Green Ⅰ可以检测 PCR 反应中获得的全部双链 DNA，但是不能区分不同的双链 DNA。为了进一步确保荧光检测的确实是靶 DNA 序列，人们又

设计了仅能与目的DNA序列特异结合的荧光探针，如TaqMan探针。TaqMan探针是一小段被设计成可以与靶DNA序列中间部位结合的单链DNA（一般为50~150 bp），并且该单链DNA的5′和3′端带有短波长和长波长两个不同荧光基团。这两个荧光基团由于距离靠近，在荧光共振能量转移(FRET)作用下会发生荧光猝灭，因而检测不到荧光。随着PCR反应的进行，TaqMan探针结合到目的DNA序列上，并且会被具有外切酶活性的Taq DNA聚合酶逐个切除而降解。被切下来的荧光基团解除了荧光猝灭的束缚，会在激发光下发出荧光，因此所产生的荧光强度直接反映了扩增靶DNA的总量(图5-10)。

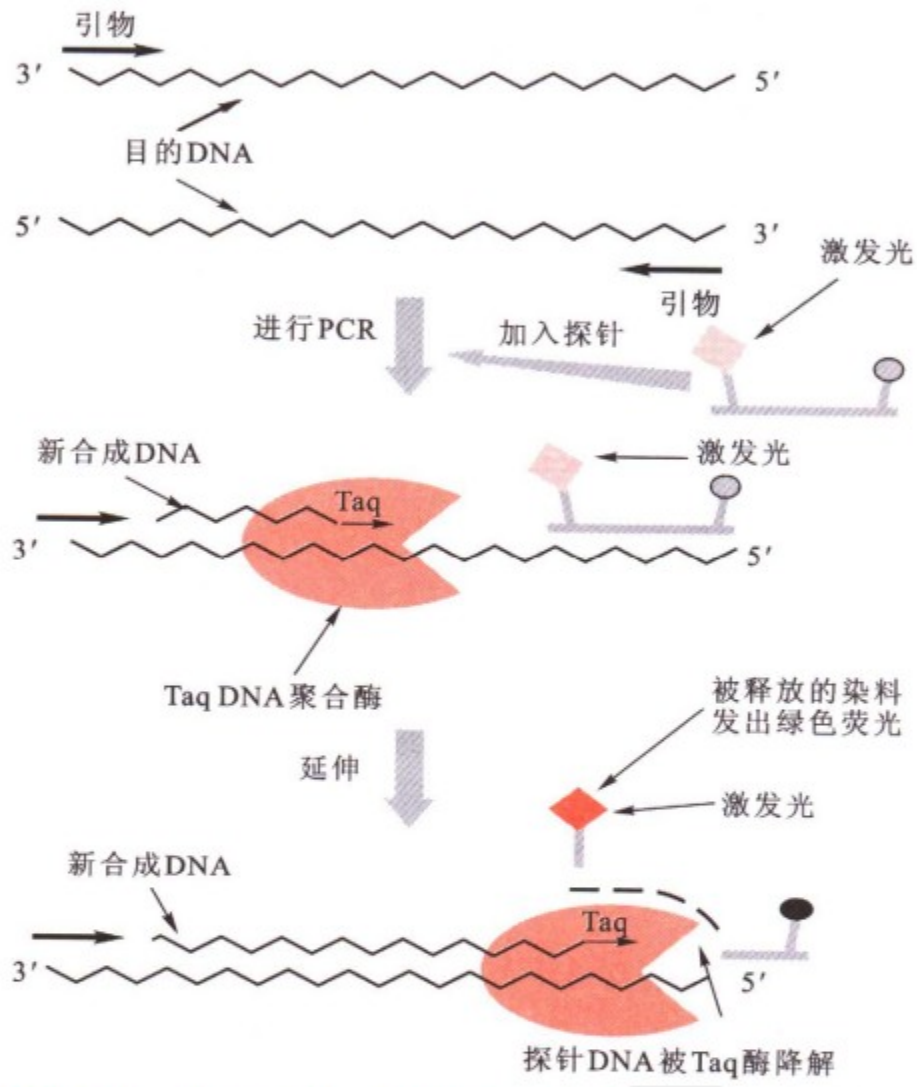



图5-10 应用TaqMan探针的实时定量PCR技术

TaqMan探针具有三个要素：一端的短波长荧光基团(菱形)，一段目的DNA特异的碱基序列和另一端的长波长荧光基团(小圆形)。探针上的荧光基团十分接近，因此两个荧光基团相互发生猝灭，不产生绿色荧光。探针上的特异碱基序列被设计成能与目的DNA中部序列结合，在PCR反应中，当Taq聚合酶延伸DNA链时，其核酸外切酶活性会逐个切掉结合在靶DNA中部的TaqMan探针序列，从而使两个荧光基团被释放出来，解除了荧光猝灭的束缚。短波长的荧光基团在激发光下产生绿色荧光，它的荧光强度反映新合成靶DNA的产量。

下面以简单的荧光染料SYBR Green Ⅰ为例，分析实时定量PCR绝对定量的结果。为确保实验数据的有效性，每次实验都应设阴性对照和4个以上标准

品(将已知浓度的 DNA 模板进行一系列的稀释后成为标准品),每个样品都至少平行做 3 个重复。通常以 10~15 个循环的荧光值作为阈值或基线,也可以以阴性对照荧光值的最高点作为基线。图 5-11a 显示随着扩增反应循环数增加,代表扩增产物含量的荧光强度增加;图 5-11b 中标准曲线由标准样品浓度的log 值和其相对应的 Ct 值所组成。Ct 值是产物荧光强度首次超过设定阈值时,PCR 反应所需的循环数。利用标准曲线,可以确定样品中待检测靶 DNA 的绝对含量(图 5-11)。此外,通过实时定量 PCR,以正常或野生型样品为参考标准,可以对处理或突变体样品中基因表达的变化多少进行相对定量,如图 5-12,表 5-3 所示。

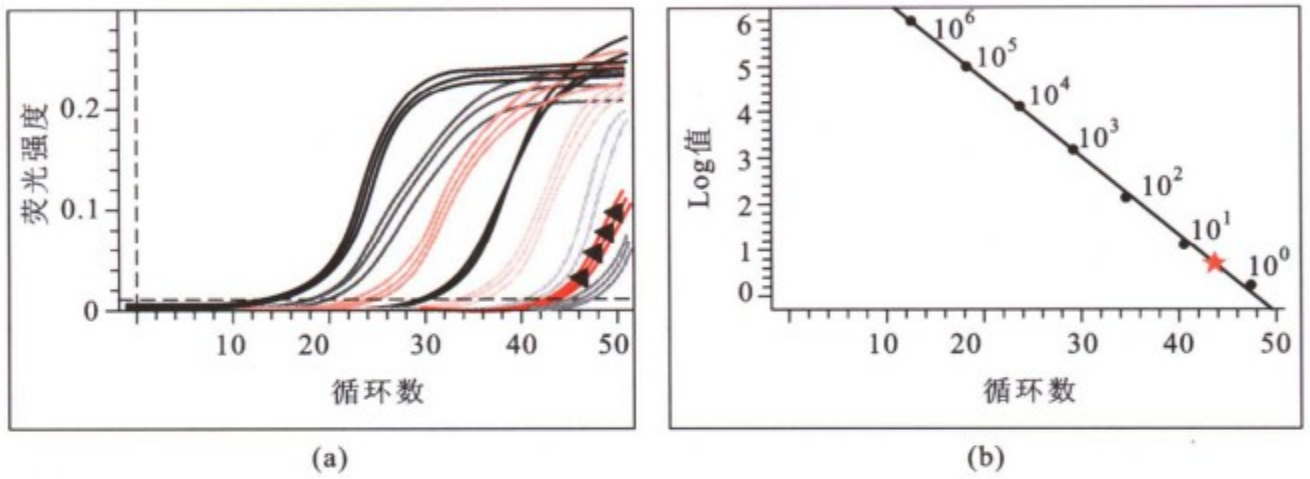



**图 5-11 用实时定量 PCR 法分析未知样品靶基因的绝对表达量** (a)扩增曲线 由已知总量(起始量从 1~1 000 000 个拷贝)的七个标准品和一个未知样品(三角形连线)得到的扩增曲线。随着 PCR 循环数(横坐标)的增加,代表 PCR 产物产量的荧光强度(纵坐标)增加。(b)标准曲线 纵坐标是每个标准品中所含 DNA 总量(log),横坐标是每个标准品 PCR 产物荧光强度首次超过设定阈值(A 中靠近底部 X 轴的虚线)时反应所需的循环数。由扩增曲线获得未知样品的 Ct 值,通过对标准曲线的线性回归分析可计算出未知样品中靶 DNA 的起始量(星号)。

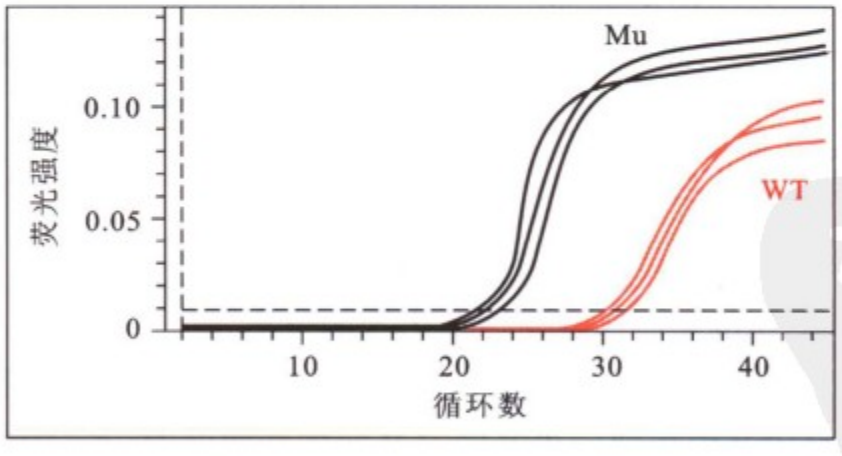



**图 5-12 实时定量 PCR 对不同样品中目的 cDNA 含量的相对定量**

野生型拟南芥(WT)整株中 *WUS* 基因的表达丰度很低,三次重复的平均 Ct 值为 31.23±0.32;而在拟南芥抑癌基因失活的突变体(Mu)整株中 *WUS* 基因的表达丰度显著增加,三次重复的平均 Ct 值为 21.92±0.13。两者的差值 ΔCt 值为 9.31±0.19,设定实验中 PCR 扩增效率为 1.8(理论值为 2),*WUS* 基因在突变体

整株与野生型拟南芥整株中表达丰度的相对比值为 $n$,$n=1.8^{\Delta Q}$。计算可得*WUS*基因在突变体整株中的表达丰度比在野生型拟南芥整株中的表达丰度提高了239±26.25 倍。数据处理见表 5-3。

表 5-3 实时定量 PCR 相对定量实验的数据处理

| | 野生型拟南芥(WT) | 突变体拟南芥(Mu) | $\triangle Ct(AvgCt_{WT}-AvgCt_{Mu})$ | 相对比值($1.8^{\triangle Ct}$) |
|---|---|---|---|---|
| *Wus* Ct 值 | 31.49 | 21.78 | | |
| | 30.87 | 21.95 | | |
| | 31.34 | 22.04 | | |
| 平均值(Avg) | 31.23 ± 0.32 | 21.92 ± 0.13 | 9.31 ± 0.19 | 239 ± 26.25 |

* 拟南芥野生型和突变体样品经由拟南芥持家基因(*UBQ*10)进行均一化处理

## 5.2.5 基因组 DNA 文库构建

把某种生物的基因组 DNA 切成适当大小,分别与载体组合,导入微生物细胞,形成克隆。汇集包含基因组中所有 DNA 序列的克隆(理论上每个序列至少有一份代表)。这样的克隆片段的总汇,称为基因组 DNA 文库。基因组 DNA 文库有十分广泛的用途,可用于分离特定的基因片段、分析特定基因结构、研究基因表达调控,还可以用于全基因组物理图谱的构建和全基因组序列测定等。

构建基因组 DNA 文库的第一步是制备大小合适的随机 DNA 片段,在体外将这些 DNA 片段与适当的载体相连成重组子,转化到大肠杆菌或其他受体细胞中,从转化子克隆群中筛选出含有靶基因的克隆。为保证能从基因组文库中筛选到某个特定基因,基因组文库必须具有一定的代表性和随机性。也就是说文库中全部克隆所携带的 DNA 片段必须覆盖整个基因组。在文库构建中通常采用两种策略提高文库代表性:一是用机械切割法或限制性内切核酸酶切割法随机断裂 DNA,以保证克隆的随机性。二是增加文库重组克隆的数目,以提高覆盖基因组的倍数。预测一个完整基因组文库应包含的克隆数目,可用 Clark 和 Carbon 于 1975 年提出的公式:

$$N=\ln(1-p)/\ln(1-f)$$

式中:$N$ 表示一个基因组文库所应该包含的重组克隆数目;

$p$ 表示所期望的靶基因在文库中出现的概率;

$f$ 表示重组克隆平均插入片段的长度和基因组 DNA 总长的比值。

为了获得整个基因簇或一个基因及其两翼延伸序列,实验中所制备的 DNA 随机片段一般约 20 kb 或更大。以人为例,其基因组大小为 $3\times10^9$ bp 若 $p=99\%$,平均插入片段大小为 20 kb,则 $N=6.9\times10^5$。

构建基因组文库最常用的是 λ 噬菌体载体(克隆能力约 15~20 kb)和限制性内切酶部分消化法。例如,用识别 4 个核苷酸的限制性内切核酸酶 *Sau*3A,它与 *Bam*H Ⅰ是一对同尾酶。因此,由 *Sau*3A 酶切产生的 DNA 片段可插入到经*Bam*H Ⅰ消化的 λ 噬菌体载体上(基本流程如图 5-13 所示)。提取真核基因组DNA,用

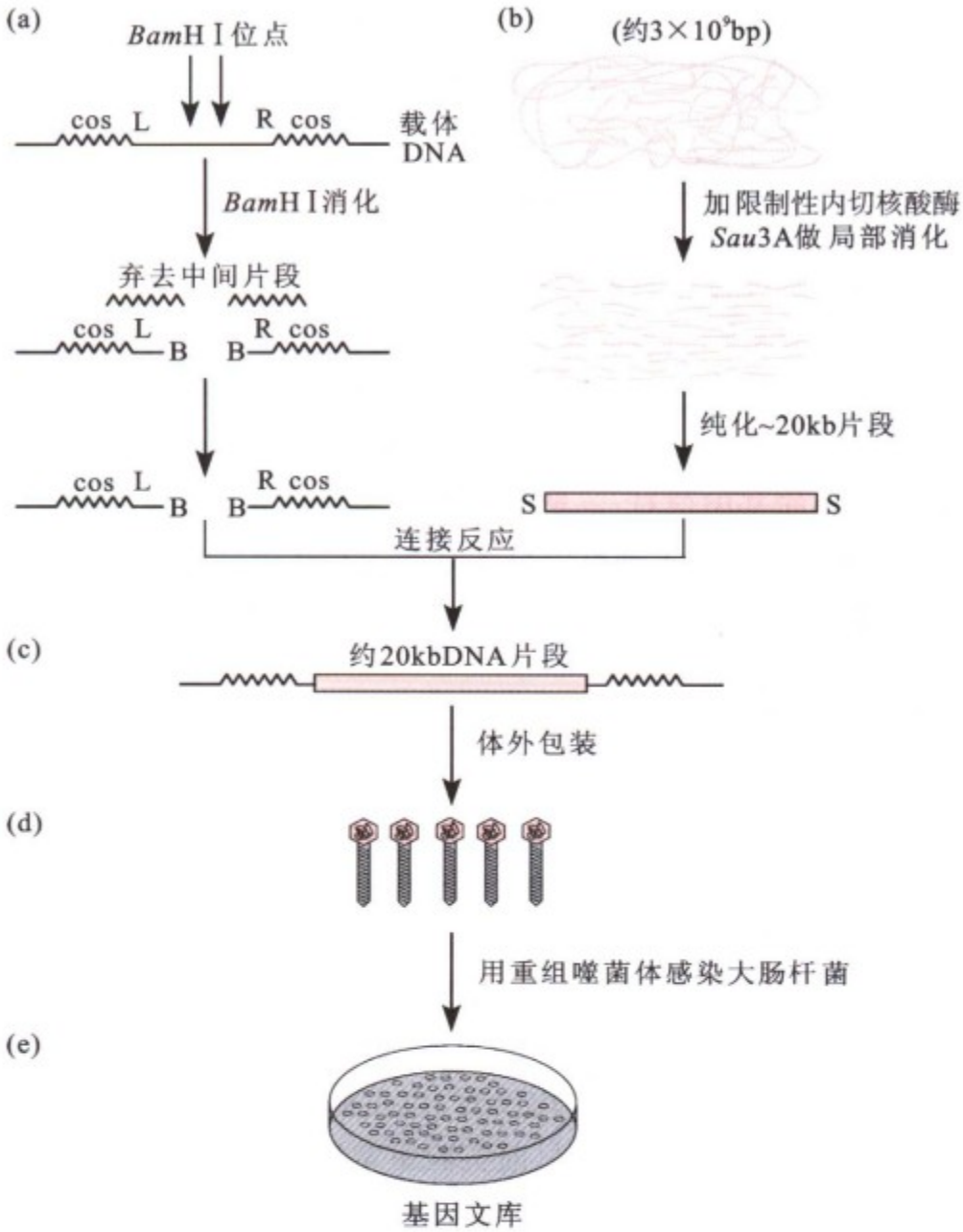



**图 5-13 用 *Sau*3A 限制性内切核酸酶消化真核生物基因组 DNA 并利用 λ 噬菌体载体构建基因组文库的过程图示** (a)载体 DNA 片段的制备;(b)真核基因组 DNA 片段的制备;(c)体外连接载体 DNA 及基因组 DNA 片段的反应;(d)体外包装;(e)感染大肠杆菌细胞。B=*Bam*H Ⅰ末端;S=*Sau*3A 末端。

*Sau*3A 局部消化,消化产物经琼脂糖凝胶电泳或者蔗糖梯度离心,收集相对分子质量为 15~20 kb 范围的 DNA 片段。同时用 *Bam*H Ⅰ消化 λ 噬菌体载体 DNA,纯化后用 T4 连接酶与收集的 DNA 片段相连接,形成嵌合分子。体外包装后用重组噬菌体感染大肠杆菌受体细胞,产生噬菌斑,组成基因组文库,它包含该真核生物基因组的绝大部分序列。λ 噬菌体文库构建方法简单高效,所获得的文库易于用分子杂交法进行筛选,因此被广泛用于细菌、真菌等基因组较小物种的研究。

除了 λ 噬菌体以外,还有很多高容量的克隆载体,如柯斯质粒、细菌人工染色体(BAC)、P1 源人工染色体(PAC)、酵母人工染色体(YAC)都可用于基因组文库的构建。这些载体的优点是可以插入大片段 DNA,如柯斯质粒可以接受大至 45 kb 的大片段的插入,应用与 λ 噬菌体文库类似,但是稳定性较差,在大量

复制的过程中容易出现缺失现象。BAC、YAC、PAC 可以容纳 70~1 000 kb 的外源片段,主要用于基因组作图、测序和克隆序列的比对。

## 5.3 RNA 基本操作技术

真核生物基因组 DNA 非常庞大,而且含有大量重复序列,无论用电泳分离技术还是用杂交方法都难以直接分离到靶基因片段。而 cDNA 则来自反转录的 mRNA,不含冗余序列,通过特异性探针筛选 cDNA 文库,可以较快地分离到相关基因。由于 RNA 分子敏感脆弱,在自然状态下难以被扩增。因此,为了研究 mRNA 所包含的功能基因信息,一般将其反转录成稳定的 DNA 双螺旋(complementary DNA,cDNA),再插入到可以自我复制的载体中。一个高质量的 cDNA 文库代表了生物体某一器官或者组织 mRNA 中所含的全部或绝大部分遗传信息。

### 5.3.1 总 RNA 的提取

细胞中的总 RNA 包括 mRNA,rRNA,tRNA 以及一些小 RNA(sRNA)。一个典型的动物细胞约含 $10^{-5}$ μg RNA,其中 80%~85%为 rRNA、15%~20%为 tRNA 及 sRNA,这些 RNA 分子具有确定的大小和核苷酸序列,特别是 rRNA 电泳后的 3 条特征性条带 28 S、18 S 和 5 S 是鉴定总 RNA 纯度和完整性的重要参数。mRNA 约占总 RNA 的 1%~5%。同时,由于 mRNA 呈单链状,容易受核酸酶的攻击,对 RNA 的操作要求比 DNA 操作更严格,操作时必须设置专门的处理方法,以保证实验环境、所用器皿及溶液均没有 RNA 酶的污染。

总 RNA 的抽提方法有多种,目前实验室常用的方法是用异硫氰酸胍-苯酚抽提法。Trizol 试剂是使用最广泛的抽提 RNA 专用试剂,主要由苯酚和异硫氰酸胍组成,可以迅速破坏细胞结构,使存在于细胞质及核内的 RNA 释放出来,并使核糖体蛋白与 RNA 分子分离,还能保证 RNA 的完整。提取 RNA 时,首先用液氮研磨材料,匀浆,加入 Trizol 试剂,进一步破碎细胞并溶解细胞成分。然后加入氯仿抽提,离心,分离水相和有机相,收集含有 RNA 的水相,通过异丙醇沉淀,可获得比较纯的总 RNA,用于下一步 mRNA 的纯化。一些公司(如 Qiagen)还设计了硅胶膜纯化柱,将含有目的 RNA 的细胞破碎液通过该柱时,RNA 吸附在硅胶膜上,从而与其他成分分开,在低盐浓度下可从硅胶膜上直接洗脱 RNA,即简化了实验程序,又得到了纯度较高的 RNA。对于某些富含酚类化合物、多糖或某些无法确定的次生代谢产物的植物组织,使用常规的 RNA 提取方法往往难以获得足够的 RNA,需要根据不同植物组织的特点,预先去除酚类物质或多糖对 RNA 的干扰。

RNA 的浓度和纯度可以通过测定其 $OD_{260}$ 和 $OD_{280}$ 来判断。$OD_{260}$ 为 1 时相当于浓度为 40 μg/mL,而 $OD_{260}/OD_{280}$ 如果在 1.8~2.0,表示所提取的 RNA 纯度较好,如果样品中有蛋白质或酚污染,则 $OD_{260}/OD_{280}$ 将明显低于 1.8。

通过琼脂糖凝胶电泳检测 RNA 与检测 DNA 的原理相类似,不同的是 RNA 呈单链状态,易形成二级结构而且易降解,因此 RNA 电泳是在变性条件下进行的。甲醛是最常用的变性剂,也可用加热方式或尿素等变性剂。聚丙烯酰胺凝胶

电泳主要用来分析小相对分子质量 RNA，电泳后如果 rRNA 大小完整，而且 28SrRNA 和 18S rRNA 亮度接近 2:1，mRNA 分布均匀，则认为 RNA 质量较好。

## 5.3.2 mRNA 的纯化

真核细胞的 mRNA 分子最显著的结构特征是具有 5′端帽子结构——$m^7G$ 和 3′端的 polyA 尾巴，这种 polyA 结构为真核生物 mRNA 的提取提供了极为方便的选择性标志，实验中常用寡聚(dT)-纤维素柱色谱法获得高纯度 mRNA。该方法利用 mRNA 3′端含有 PolyA⁺的特点，当 RNA 流经寡聚 dT 纤维素柱时，在高盐缓冲液的作用下，mRNA 被特异性地结合在柱上，再用低盐溶液或蒸馏水洗脱 mRNA。经过两次寡聚 dT 纤维柱后可得到较高纯度的 mRNA。目前有许多商业化试剂盒可用于 mRNA 的纯化，如 Promega 公司的 polyAT Tract mRNA 分离系统将生物素标记的寡聚 dT 引物与细胞总 RNA 温育，加入与微磁球相连的抗生物素蛋白以结合 polyA mRNA，通过磁场吸附作用将 polyA mRNA 从总 RNA 中分离(图 5-14)。

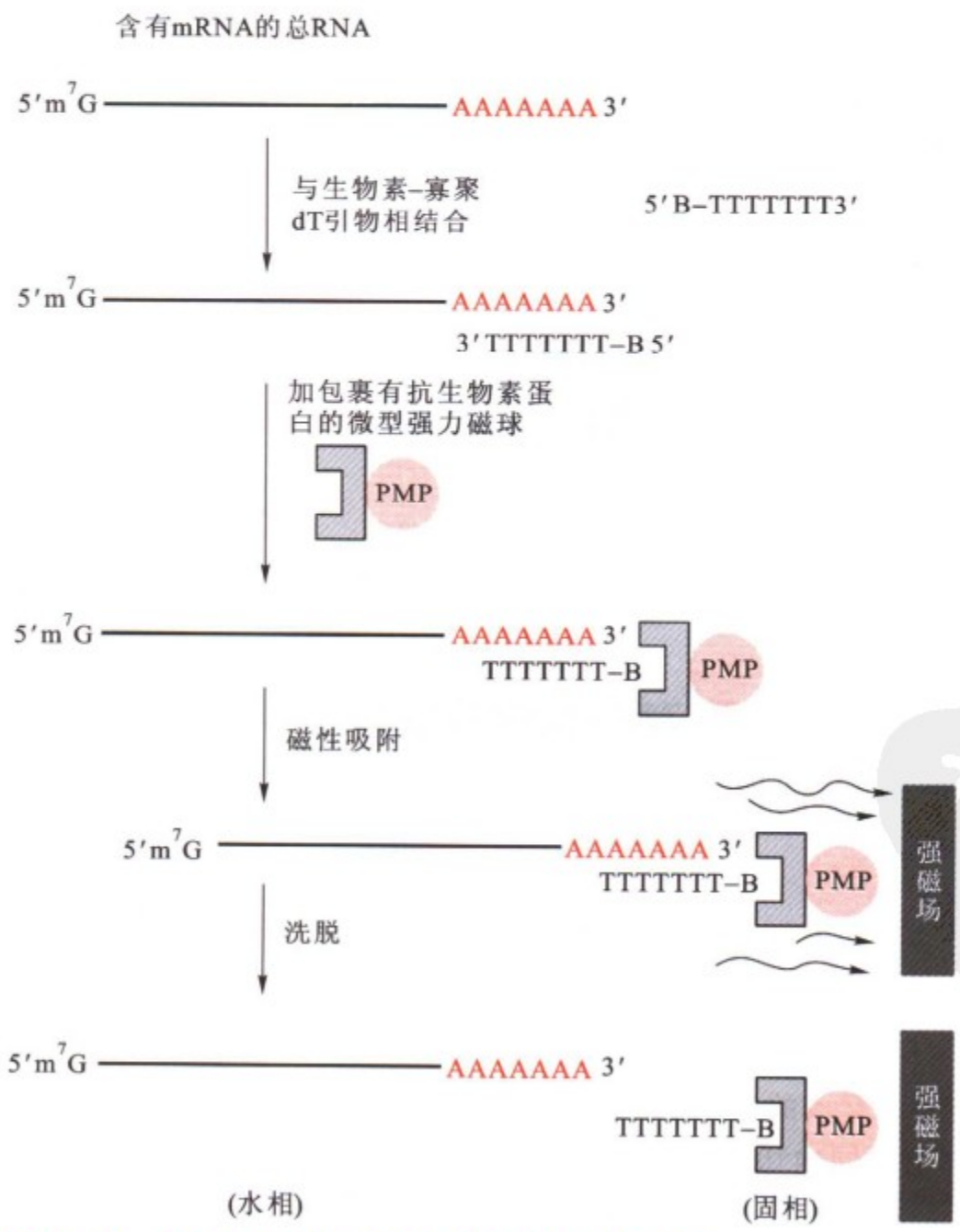



图 5-14 PolyAT Tract mRNA 的分离纯化过程简图

### 5.3.3 cDNA 的合成

cDNA 的合成包括第一链和第二链 cDNA 的合成,整个流程如图(图 5-15)所示。第一链 cDNA 的合成是以 mRNA 为模板,反转录成 cDNA,由反转录酶催化,该酶合成 DNA 时需要引物引导,常用的引物是 oligo dT。Oligo dT 引物一般包含 12~20 个脱氧胸腺嘧啶核苷酸,后面加一个连接引物(通常为 *Xho* Ⅰ 等酶切位点)以便于克隆构建。

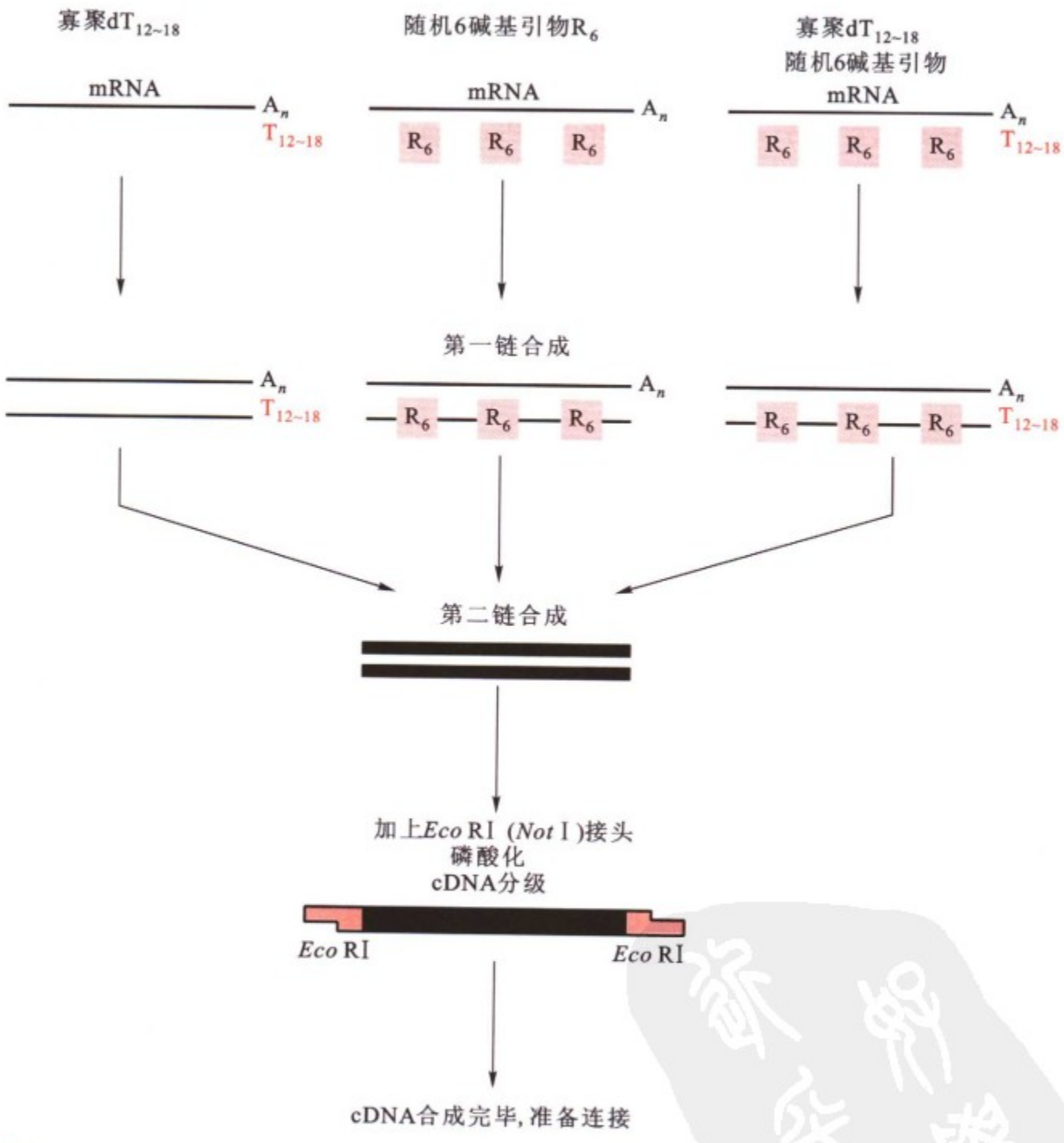
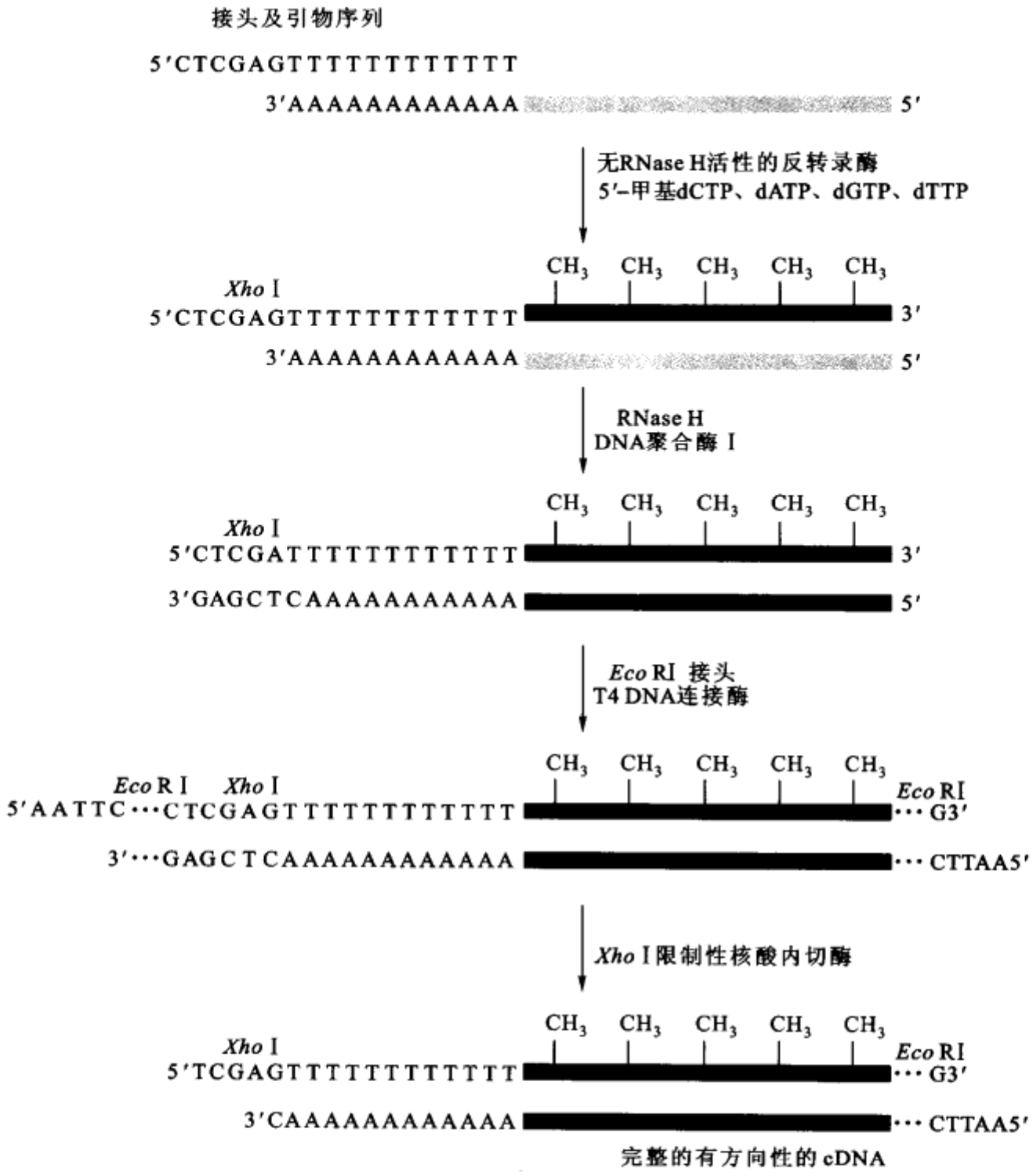



图 5-15 cDNA 合成过程示意图

在 cDNA 合成的过程中,应选用活性较高的反转录酶及甲基化 dCTP,cDNA 两端应加上不同内切酶所识别的接头序列,保证所获得双链 cDNA 的方向性(图 5-16)。反应体系中一般加入甲基化 dCTP,保证新合成的 cDNA 链被甲基化修饰,以防止构建克隆时被限制性内切酶切割。第二链 cDNA 的合成是以第一




▌图 5-16 定向 cDNA 合成及分子修饰

链为模板,由 DNA 聚合酶催化。常用 RNase H 切割 mRNA-cDNA 杂合链中的 mRNA 序列所产生的小片段为引物合成第二条 cDNA 的片段，再通过 DNA 连接酶的作用连成完整的 DNA 链。此时加入含有另一个酶切位点的黏性接头(如 *Eco* RI),与 cDNA 相连接后用 *Xho* I 酶切,使 cDNA 双链 5′端和 3′端分别具有 *Eco* RI和 *Xho* I 黏性末端,保证它与载体相连时有方向性。因为绝大多数大肠杆菌细胞都会切除带有 5′-甲基胞嘧啶的外源 DNA，所以实验中常选用mcrA⁻ mcrB⁻菌株以防止 cDNA 被降解。

### 5.3.4 cDNA 文库的构建

由于 cDNA 的长度一般在 0.5~8 kb,常用的质粒载体和噬菌体类载体都能

满足要求。cDNA 文库的载体选择要根据该文库的用途来确定，例如，常用的 Uni-zap XR 载体是一种 λ 噬菌粒载体，具备噬菌体的高效性和质粒载体系统可利用蓝白斑筛选的便利，可容纳 0~10 kb DNA 插入片段。该载体内部含有 pBluescript 载体的全部序列，重组后可通过体内剪切反应（*in vivo* excision）将 cDNA 插入片段转移到质粒系统中进行筛选、克隆和序列分析。

含有 cDNA 插入片段的重组噬菌粒只有经过体外蛋白外壳包装反应，才能成为有侵染和复制能力的成熟噬菌体。包装反应的实验条件要求非常精准，DNA 和蛋白提取物的浓度、两者的体积比、反应温度及时间对包装效果都有很大影响，而包装反应结果的好坏对重组噬菌体的侵染能力有重要影响。用经包装的噬菌体感染大肠杆菌培养物后涂布在琼脂平板上，便会产生成千上万个独立噬菌斑，每个噬菌斑由一个重组噬菌体分子形成。一个比较完整的 cDNA 文库常包含大于 $5\times10^5$ 的独立克隆。一旦获得含有某种组织器官 cDNA 信息的噬菌粒文库，就可用于筛选目的基因、大规模测序、基因芯片杂交等功能基因组学研究。

## 5.3.5 基因文库的筛选

基因文库的筛选是指通过某种特殊方法从基因文库中鉴定出含有所需重组 DNA 分子的特定克隆的过程。筛选的方法有很多种，如核酸杂交法，PCR 筛选法和免疫筛选法等。

### 5.3.5.1 核酸杂交法

核酸杂交法以其广泛的适用性和快速性成为基因文库筛选中最常用的一种方法，用放射性标记的特异 DNA 探针适用于高密度的菌落杂交筛选（图 5-17）。将一张圆形硝酸纤维素膜放在一个装有琼脂培养基的培养皿表面，将待筛选菌落从其生长的平板上转移到硝酸纤维素膜上，进行适当温育。同时保留原来的菌落平板做对照。取出已经长有菌落的膜，用碱液处理，使菌落发生裂解，

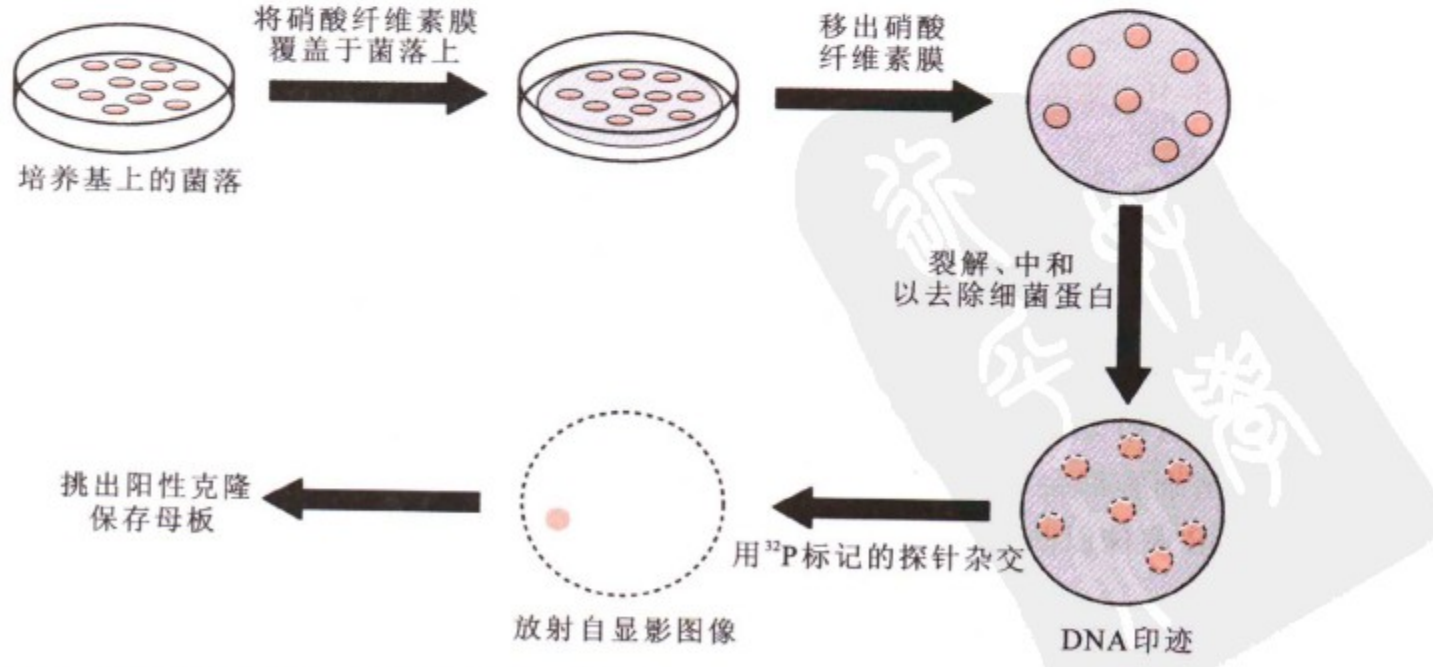



图 5-17 通过核酸杂交筛选目的克隆流程图

DNA 随之变性。接着用蛋白酶 K 处理硝酸纤维素膜去除蛋白质,形成菌落 DNA 的印迹。80℃烘烤滤膜,将 DNA 固定在膜上。将滤膜与放射性标记的 DNA 或 RNA 探针杂交,通过放射性自显影显示杂交结果。在 X 线片上显黑色斑点的就是实验中寻找的目的克隆,可以通过对应于平板上的位置找到相应克隆。

与菌落杂交相似,也可用杂交筛选法进行重组噬菌斑的筛选。将硝酸纤维素膜覆盖在琼脂平板表面,使之与噬菌斑直接接触,噬菌斑中大量没有被包装的游离 DNA 及噬菌体颗粒便一同转移到膜上。可从同一噬菌斑平板上连续印几张同样的硝酸纤维素膜,进行重复实验,而且可以使用两种或数种不同的探针筛选同一套重组子,提高结果的可靠性。

5.3.5.2 PCR 筛选法

PCR 筛选法与核酸杂交筛选法具有同样的通用性,而且操作简单,但前提是已知足够的序列信息并获得基因特异性引物。例如,要从一个基因组 DNA 文库筛选目的基因,首先将整个文库(以质粒的形式或者细菌的形式均可)保存在多孔培养板上,用设计好的目的基因探针对每个孔进行 PCR 筛选,鉴定出阳性的孔,把每个阳性孔中的克隆再稀释到次级多孔板中进行 PCR 筛选。重复以上程序,直到鉴定出与目的基因对应的单个克隆为止。

5.3.5.3 免疫筛选法

由于免疫筛选法是基于抗原抗体特异性结合原理,所以该法适用于对表达文库的筛选。也就是说 DNA 文库或者 cDNA 文库是用表达载体构建的,每个克隆都可以在宿主菌中表达,产生一段多肽。如果实验中靶基因的序列完全未知但是有针对该基因产物的特异性抗体,就可以用免疫筛选法进行筛选。

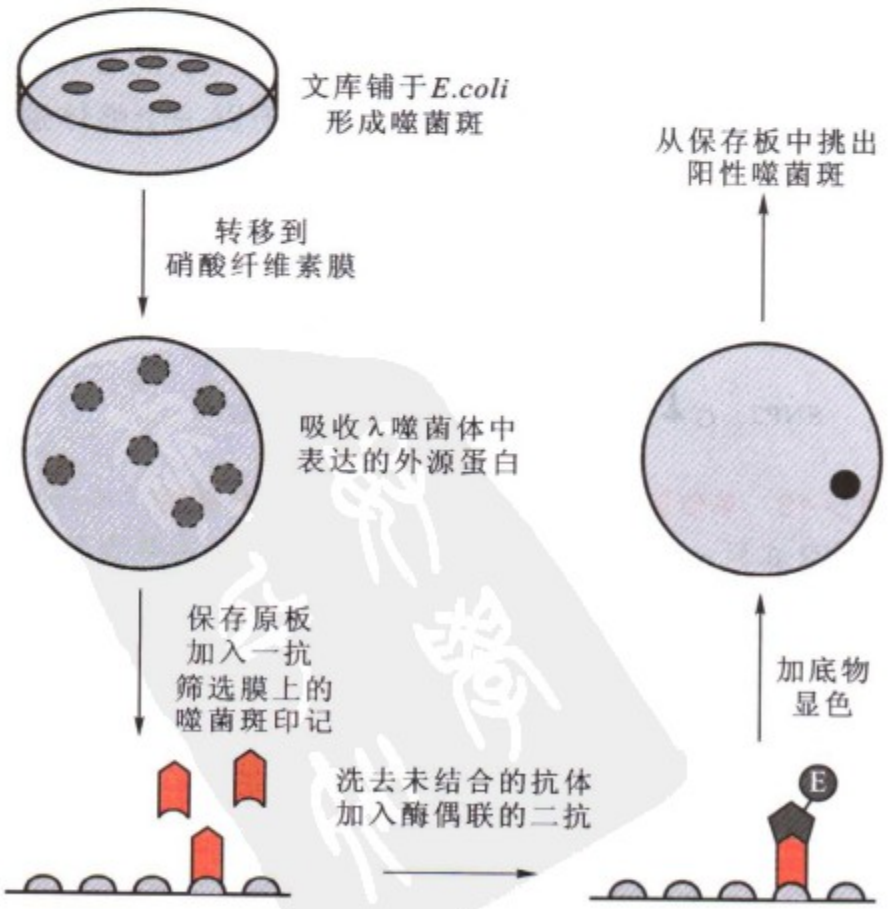



图 5-18 噬菌体表达文库的免疫化学筛选法示意图

免疫检测与菌落或噬菌斑的核酸杂交相似,先将菌落或噬菌斑影印到硝酸纤维素膜上,原位溶解菌落释放抗原蛋白,再用抗体与固定了抗原的膜杂交,抗原抗体结合后,再用标记的二抗与之反应,通过对标志物的检测,就可以找到阳性克隆(图 5-18)。

## 5.4 SNP 的理论与应用

SNP 是 single nucleotide polymorphism 的缩写,中文翻译为单核苷酸多态性,指基因组 DNA 序列中由于单个核苷酸(A,T,C 和 G)的突变而引起的多态性。染色体 DNA 同一位置上的每个碱基类型叫做一个等位位点。SNP 是基因组中最简单、最常见的多态性形式,具有很高的遗传稳定性。随着 SNP 检测和分析技术的进一步发展,已经成为继限制性片段长度多态性(RFLP)和微卫星标记(SSR)之后的第三代遗传标记。

### 5.4.1 SNP 概述

一个 SNP 表示在基因组某个位点上一个核苷酸的变化,这种变化可能是转换(C⟷T,在其互补链上则为 G⟷A),也可能是颠换(C⟷A,G⟷T,C⟷G,A⟷T),具有转换型变异的 SNP 约占 SNP 总量的 2/3 左右。转换的概率之所以高,是因为 CpG 二核苷酸上的胞嘧啶(C)残基是人类基因组中最易发生突变的位点,其中大多数是甲基化的,可自发地脱去氨基而形成胸腺嘧啶(T)。SNP 广泛存在于人类基因组中,其发生频率约为 1%或更高。据估计,人类 DNA 中每 300~1 000 bp 就有一个 SNP,因此,一个人类个体大约携带 300 万~1 000 万个 SNPs。位于染色体上某一区域的一组相关联的 SNP 等位位点被称作单倍型(haplotype),相邻 SNPs 的等位位点倾向于以一个整体遗传给后代(图 5-19)。

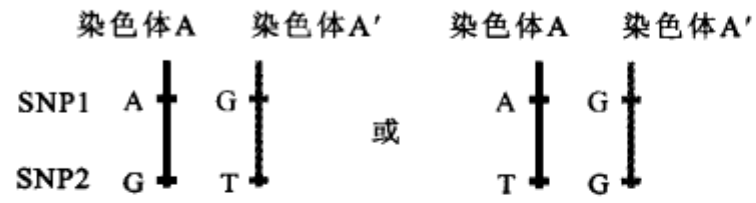



**图 5-19 单倍型是染色体上共同遗传的 SNP 组合** 注:SNP1 和 SNP2 在同一条染色体上而且距离很近,SNP1 有等位位点 A 和 G,SNP2 有等位位点 G 和 T。因此,染色体对有两种可能的 SNP 组合。

根据 SNP 在基因组中的分布位置可分为基因编码区 SNP(cSNP),基因调控区 SNP(pSNP)和基因间随机非编码区 SNP(rSNP)等三类。因为编码区内的变异率仅占周围序列的 1/5,cSNP 的总量显著少于其他两类 SNPs。从对生物遗传性状的影响上看,cSNP 又可分为 2 种,一种是同义 cSNP (synonymous cSNP),即 SNP 所导致的编码序列改变并不影响其所翻译的蛋白质的氨基酸序列,突变碱

基与未突变碱基的含义相同;另一种是非同义 cSNP(non-synonymous cSNP),指碱基序列的改变可使以其为蓝本翻译的蛋白质序列发生改变,从而影响蛋白质的功能。这种改变常常是导致生物性状改变的直接原因。位于基因调控区的 SNP 则会影响基因表达量的多少,因此,这两类 SNP 在功能和疾病发生发展方面具有更重要的意义。

研究表明,某玉米储藏蛋白的 8 个基因型共包含了 3 个单倍型 (H1,H2 和 H3),它们分别由多个相邻的 SNP 所构成。在 H1 和 H2 中,外显子Ⅳ的 57 位(C→T)、内含子Ⅳ的 232 位和 236 位(C→T),共有 3 个 SNPs,说明这两个单倍型至少是 3 个等位基因的集合。在 H3 和 H2 中,外显子Ⅳ的 130 位(T→C)、165 位(A→T),内含子Ⅳ的 232 位(C→T)、236 位(C→T),共有 4 个 SNPs(表5-4)。

表 5-4 玉米储藏蛋白基因 globulin1 不同单倍型和基因型之间的多态性比较

| 单倍型 | 基因型 | 外显子Ⅳ | | | 内含子Ⅳ | |
|---|---|---|---|---|---|---|
| | | 57 | 130 | 165 | 232 | 236 |
| H1 | H-4 | T | T | A | T | T |
| | H-3 | T | T | A | T | T |
| | H-1 | T | T | A | T | T |
| | E-2 | T | T | A | T | T |
| | B-73 | T | T | A | T | T |
| H2 | MO17 | C | T | A | C | C |
| | B-1 | C | T | A | C | C |
| H3 | H60 | C | C | T | T | T |

## 5.4.2 SNP 的检测技术

传统的 SNP 检测方法是采用已有技术,如限制性酶切片段长度多态性(RFLP)、PCR-单链构象多态性(PCR-SSCP)、毛细管电泳及变性高效液相色谱(DHPLC)等。由于上述技术都必须通过凝胶电泳进行分析,通量受到限制。而且这些方法仅能判断 SNP 的有无而不能知道确切的碱基类型,只有进行 DNA 序列分析才能确认所发现的 SNP。因此,目前国际上最常见的仍然是通过 DNA 测序法获得新的 SNP。

基因分型(genotyping)是指利用数据库中已有的 SNP 进行特定人群的序列和发生频率的研究,主要包括基因芯片技术、Taqman 技术、分子信标(molecular beacon)技术和焦磷酸测序法(pyrosequencing)等。

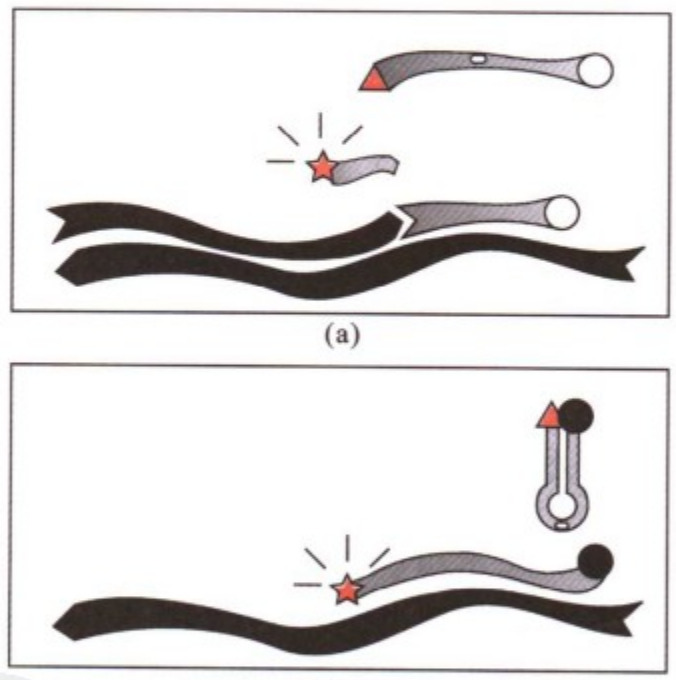
1. 基因芯片技术

基因芯片技术是在固相支持介质上进行分子杂交和原位荧光检测的一种高通量 SNP 分析方法。通过优化芯片杂交程序,使探针只与完全互补的序列杂交而不与含有单个错配碱基的序列杂交。待测基因样品经 PCR 扩增及荧光化学标记后与固定的探针进行杂交。由于目标基因和探针杂交的程度与荧光强度

及种类相关,因此通过激光扫描,可根据荧光强弱或荧光的种类检测出被检序列的碱基类别。美国昂菲公司已制成此类芯片,其优点是高通量,一次可对多个SNP 进行规模性筛选,被检起始材料也很少,操作步骤简单。缺点是芯片设计成本高,由于 DNA 样品的复杂性,有些 SNP 不能被检出。

2. Taqman 技术

在 PCR 反应系统中加入 2 种不同荧光标记的探针,它们分别与两个等位基因完全配对。探针的设计运用了荧光共振能量转换(fluorescence resonance energy transfer,FRET)技术,探针的 5′端和 3′端分别用特殊染料标记,称为供体–受体染料对或引爆–猝灭染料对。正常情况下,由于探针 5′端荧光基团和 3′端猝灭基团紧邻在一起,供体所发荧光由于其光谱在空间上接近受体染料而被猝灭,只有当两者分离时才能检测到供体所发出的荧光。随着 PCR 的有效进行,与模板完全配对的探针逐步被 Taq DNA 聚合酶的 5′→3′外切活性所切割,致使探针 5′端上的荧光基团与 3′端的猝灭基团分离,猝灭效应解除,供体的荧光基团被激活(图 5–20a);而与模板不能完全配对的探针(代表另一种等位基因)不能被有效切割,供体荧光基团依然被猝灭。通过相应仪器检测荧光值的变化,便可进行基因分型。



(a)

(b)

**图 5–20 Taqman 技术及分子信标技术原理示意图** (a)Taqman 技术示意图;(b)分子信标技术示意图。黑色条带表示待检测 DNA 序列,灰色短条带表示探针,中间有白色小格表示该探针与待测序列不完全匹配,三角和圆圈表示供体和受体染料对,星形表示激发荧光。

3. 分子信标技术

与 Taqman 技术相似,分子信标技术(molecular beacon)也是在 PCR 反应体系中加入荧光标记的探针与靶序列杂交,通过仪器检测荧光值的变化,进行基因分型。分子信标是一种呈茎环结构的荧光标记的寡核苷酸,环状区与靶序列特异性结合,茎干区由 5~8 个碱基对组成,5′端带有荧光发生基团,3′端带有荧

光猝灭基团。自由状态时,分子信标呈发卡式结构,从而荧光基团和猝灭基团相距较近(约 7~10 nm),发生荧光共振能量转移,使荧光基团发出的荧光被猝灭基团吸收并以热的形式散发,荧光几乎完全被猝灭。当分子信标与靶分子相结合时,荧光基团和猝灭基团之间的距离增大,分子信标的荧光几乎 100%恢复,所检测到的荧光强度与溶液中靶标量成正比(图 5-20b)。如果靶序列中存在错配碱基,分子信标与靶序列的结合就不稳定,就不会产生荧光。分子信标法在选择荧光染料时,不必像 Taqman 法那样考虑染料之间光谱的重叠性,一次可以使用 4 种或 4 种以上染料,同时对多个 SNP 进行分析。Taqman 法和分子信标法优点都是操作简单,可以自动化;缺点是不能达到高通量分析,荧光探针费用高。

4. 焦磷酸测序法

焦磷酸测序法(pyrosequencing)是一种不依赖平板胶或毛细管电泳,不依赖 DNA 的荧光标记/激发/检测体系的序列分析技术,适用于已知序列的 DNA 片段进行验证分析,适用于已知 SNP 的序列验证及基因分型。焦磷酸测序法主要是由四种酶催化同一反应体系中的酶级联反应,包括 DNA 聚合酶 (DNA polymerase)、硫酸化酶 (ATP sulfurylase)、萤光素酶 (luciferase) 和双磷酸酶 (apyrase),反应底物为 adenosine 5′phosphosulfate(APS)和萤光素(luciferin)。反应体系还包括待测序的 DNA 单链和测序引物。在每一轮测序反应中,加入一种 dNTP,如该 dNTP 与模板配对,聚合酶就可以催化该 dNTP 掺入到引物链中并释放焦磷酸基团 (PPi)。掺入的 dNTP 和释放的焦磷酸的物质的量相等。反应时 dATP 由 deoxyadenosine alfa-thio triphosphate(dATPαS)替代,因为 DNA 聚合酶对 dATPαS 的催化效率比对 dATP 的催化效率高,且 dATP 是萤光素酶的底物,dATPαS 不是。硫酸化酶催化 APS 和 $PP_i$ 形成 ATP,其物质的量和焦磷酸一致。ATP 驱动萤光素酶介导的萤光素向氧化萤光素(oxyluciferin)的转化,氧化萤光素发出与 ATP 含量成正比的可见光信号。光信号由 CCD 摄像机检测并通过相应的软件得到反映。每个峰的高度(光信号)与反应中掺入的核苷酸数成正比。ATP 和未掺入的 dNTP 由双磷酸酶降解,猝灭光信号,并再生反应体系,加入另一种 dNTP 进行下一轮反应。随着以上过程的循环进行,互补 DNA 链合成,DNA 序列信号峰确定(图 5-21)。该技术分析系统自动化程度高,通量大,速度快,易于建立标准化操作,适合大规模 SNP 研究及基因分型。

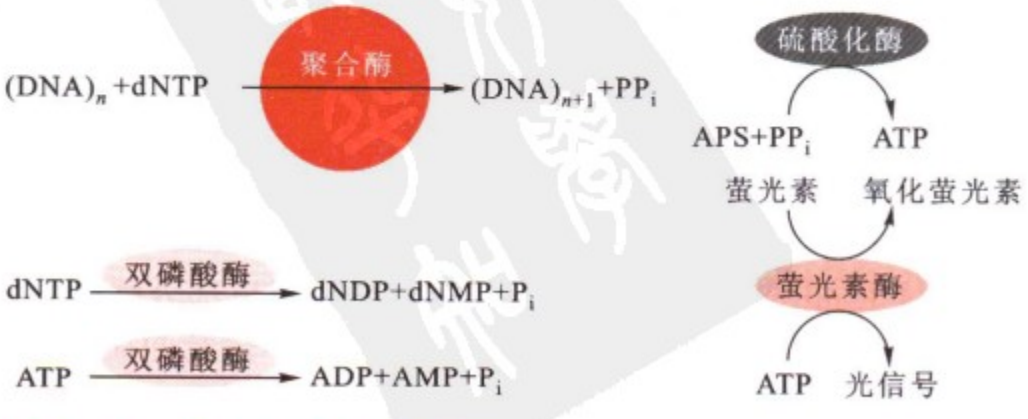



图 5-21 焦磷酸测序法示意图

### 5.4.3 SNP 的应用

1. 人类基因单体型图的绘制

大多数染色体区域只有少数几个常见的单体型,代表了一个群体中人与人之间的大部分多态性。某些染色体区域可以有很多 SNP 位点,但是只用少数几个标签 SNPs,就能够提供该区域内大多数的遗传多态性模式。绘制人类基因单体型图的目的就是描述人类常见的遗传多态性模式和染色体上具有成组紧密关联 SNPs 的区域。2002 年 10 月,国际人类基因组单体型图计划(HapMap 计划)正式启动,截至目前已有超过 150 万 SNPs 被精确定位于各染色体上,并且已经根据这些 SNPs 对来自 4 个不同人种的 269 份 DNA 样品进行了基因分型。后期目标是要使总密度达到每 500 bp 有一个 SNP。通过不同个体基因组 DNA 的基因分型和频率计算,绘制更加精密的单倍型图谱。人类基因单体型图的绘制完成,将为我们精确定位复杂疾病(如糖尿病,癌症,心脏病,脑卒中和哮喘等)易感基因提供重要信息。

2. SNP 与疾病易感基因的相关性分析

当一个遗传标记的频率在患者中明显超过非患者时,就表明该标记可能与这种疾病相关。随着大量代谢通路和上百万 SNPs 的确认,SNP 作为新一代遗传标记在人类疾病研究中显示出极高的潜在价值。已经有高血压、哮喘、类风湿关节炎、肺癌、前列腺癌等许多易感基因通过 SNP 的相关研究被发现。

3. 指导用药与药物设计

由于 SNP 能够充分反映个体间的遗传差异,所以通过研究 SNP 与个体对药物敏感或耐受的相关性研究,可能阐明遗传因素对药效的影响,因此可能建立与基因型相关的治疗方案,对患者施行个性化用药。另外,随着 SNP 的研究与药物基因组学的结合,根据特定的基因型来设计药物将成为可能。

## 5.5 基因克隆技术

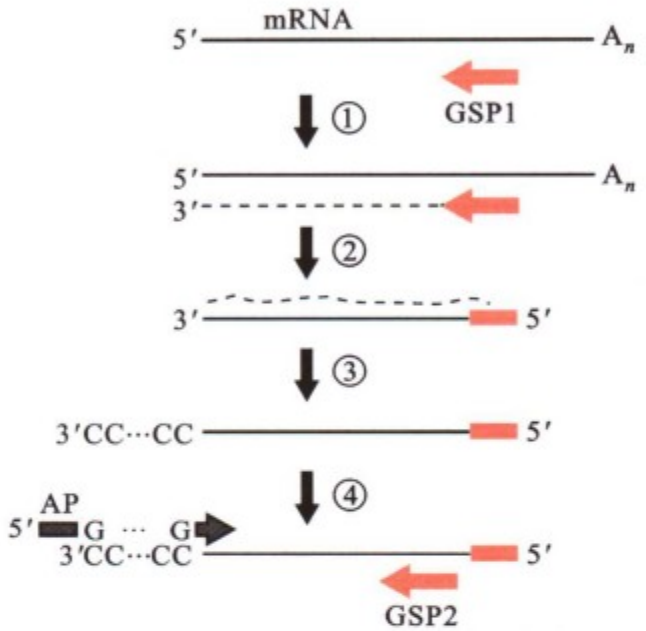
在当今生命科学的各个研究领域中,“克隆”(clone)一词已被广泛使用。在多细胞的高等生物个体水平上,人们用克隆表示由具有相同基因型的同一物种的两个或数个个体组成的群体,所以说,从同一受精卵分裂而来的单卵双生子(monozygotic twins)便是属于同一克隆。在细胞水平上,克隆一词是指由同一个祖细胞(progenitor cell)分裂而来的一群带有完全相同遗传物质的子细胞。在分子生物学中,人们把将外源 DNA 插入具有复制能力的载体 DNA 中,使之得以永久保存和复制的过程称为克隆。

真核生物基因组 DNA 十分庞大,而且含有大量的重复序列。因此,无论是采用电泳分离技术还是通过杂交的方法,都难以直接分离到目的基因片段。这个问题可以通过由 mRNA 产生的 cDNA 进行克隆而得以部分解决,因为尽管高等生物一般具有约 3 万~5 万种不同的基因,但在一定时间阶段的单个细胞或组织中,仅有 15%左右的基因得以表达,产生出约 5 000~10 000 种不同的 mRNA 分子。CDNA 克隆的基本过程是通过一系列酶催作用,使总 poly A mRNA

转变成双链 cDNA 群体并插入到适当的载体分子上，转化大肠杆菌宿主菌株细胞，构成包含所有基因编码序列的 cDNA 基因文库。

## 5.5.1 RACE 技术

RACE(rapid amplification of cDNA ends)技术是一项在已知 cDNA 序列的基础上克隆 5′端或 3′端缺失序列的技术，在很大程度上依赖于 RNA 连接酶连接和寡聚帽子的快速扩增。已知 cDNA 序列可来自序列表达标签、减法 cDNA 库、差式显示和基因文库筛选。主要操作步骤如下(图 5-22)。

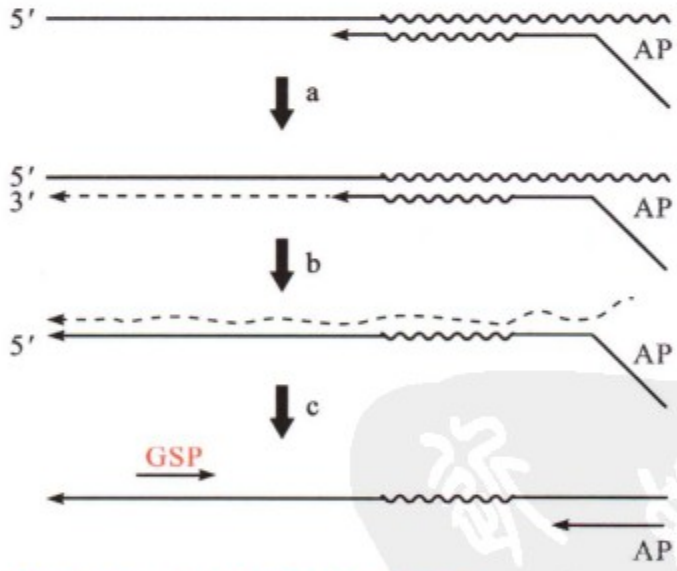



①在反转录酶的作用下，用已知基因片段内部特异引物(gene-specific primer 1，GSP1)起始 cDNA 第一条链的合成

②用 RNase 混合物降解模板 mRNA 并纯化已合成的 cDNA 第一条链

③由末端转移酶在已合成的 cDNA 链的 cDNA 链 3′端连续加入 dCTP，形成 oligo dC 尾巴

④以连有 oligo dC 的锚定引物(anchor primer，AP) 和基因片段内部特异引物 GSP2 进行 nest PCR 扩增，以期得到目的基因 5′端片段并检测




a. 在反转录酶的作用下，以连有可以与 mRNA 3′多核苷酸尾配对的 oligo dT 的锚定引物 AP 起始 cDNA 第一条链的合成

b. 用 RNase 降解模板 mRNA

c. 用锚定引物 AP 和基因片段内部特异引物 GSP 进行 PCR 扩增，以期得到目的基因 3′端片段，并进行检测

图 5-22 5′和 3′RACE 流程图

1. 获得高质量总 RNA。包括完整的 mRNA、tRNA、rRNA 和部分不完整的 mRNA。

2. 去磷酸化作用。带帽子结构的 mRNA 不受影响。

3. 去掉 mRNA 的 5′帽子结构，加特异性 RNA 寡聚接头并以 RNA 连接酶相连接。

4. 以特异性寡聚 dT 为引物，在反转录酶的作用下，反转录合成第一条

cDNA 链,包含了寡聚接头的互补序列。

5. 分别以第一链 cDNA 为模板进行 RACE 反应。5′RACE 以 5′端 RNA 寡聚接头的部分序列和基因特异的 3′端反向引物进行 PCR 扩增，获得基因的 5′端序列。3′RACE 以 5′端基因特异的引物和 3′端寡聚 dT 下游部分序列为引物进行 PCR 扩增,获得基因的 3′端序列。如果只对 3′端序列感兴趣,可以越过 2、3 步,直接从第 4 步开始进行 3′RACE。

6. 将纯化后的 PCR 产物克隆到载体 DNA 中,进行序列分析。除了获得全长 cDNA 之外,RACE 技术还被用于识别 5′和 3′端非转录序列，研究转录起始位点的不均一性,研究启动子区的保守性。

## 5.5.2 应用 cDNA 差示分析法克隆基因

尽管分子生物学研究数十年来得到了长足的发展，大量分子克隆技术日臻成熟和完善,但是,要在没有任何探针的情况下克隆控制某一特定性状或生理反应中间步骤的基因仍然有很强的挑战性。作为对染色体步移技术、定位克隆技术和

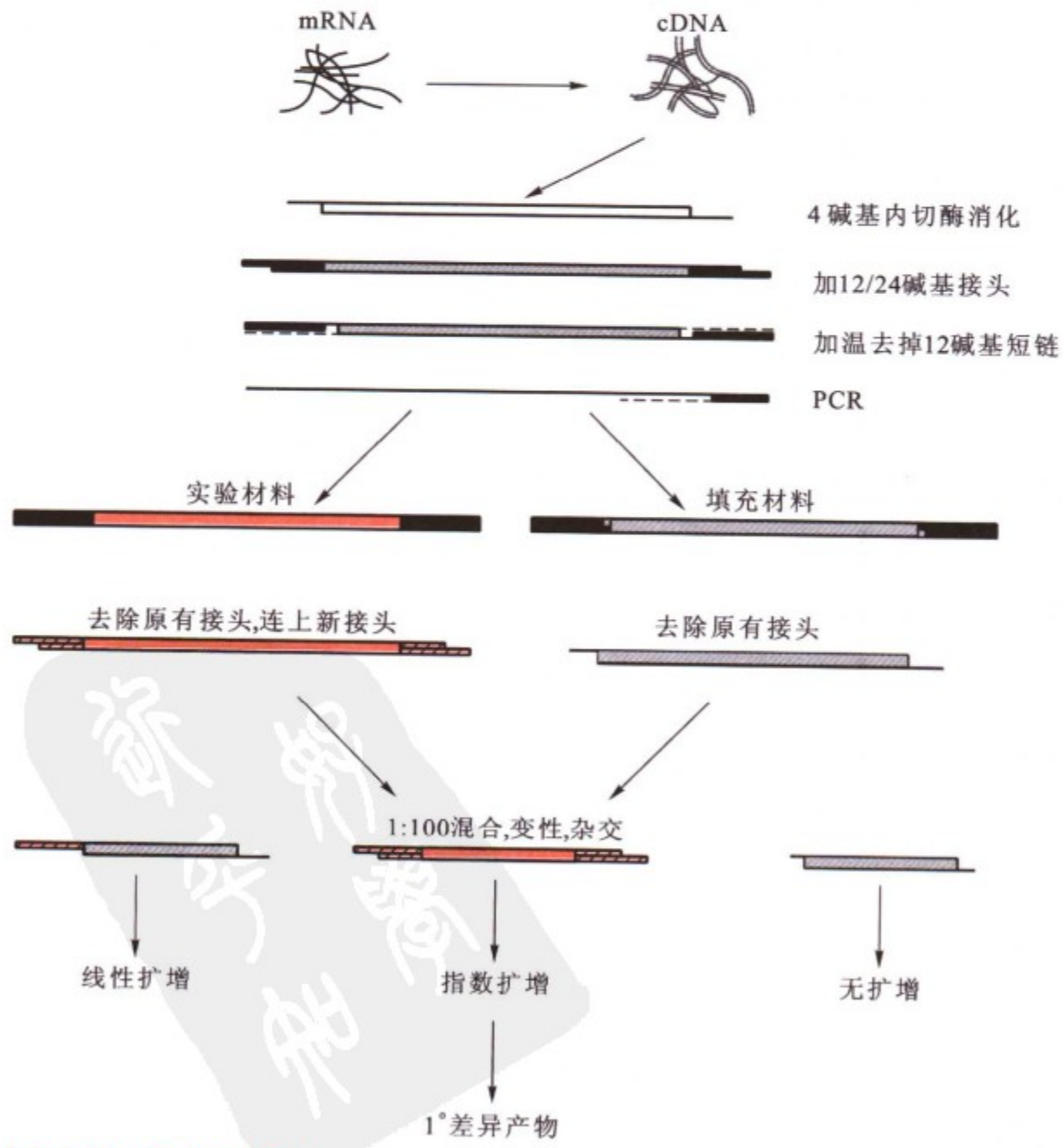



图 5-23 RDA 流程图

cDNA 差显技术等方法的补充,cDNA 差示分析法(representational difference analysis,RDA)充分发挥了 PCR 能以指数形式扩增双链 DNA 模板,而仅以线性形式扩增单链模板的特性,通过降低 cDNA 群体复杂性和更换 cDNA 两端接头等方法,特异性扩增目的基因片段(图 5-23)。因为试验材料(tester)和探针材料(driver)在接受差示分析前均经一个 4 碱基切割酶处理,形成平均长度256bp 的代表群(representation),保证绝大部分遗传信息能被 PCR 扩增。第一次T 减 D 反应中两者的物质的量比就要求达到 1:100 或更高, 经过 2~3 次重复后,T 群体中非特异性序列几乎没有偶然逃脱的可能性。每次 T 减 D 反应后仅设置72℃复性与延伸,94℃变性这两个参数共 20 个 PCR 循环,PCR 产物的特异性和所得 cDNA 片段的纯度均较高。

### 5.5.3 Gateway 大规模克隆技术

随着越来越多的动植物基因组全序列被测定,前所未有的编码基因(可译框架)被识别,但是,目前还不清楚绝大部分基因的功能。要阐明这些基因的功能,就必须把不同的可译框架连入各种载体,进行蛋白质表达、抗体制备、宿主细胞或植物转化、表型分析、细胞内定位以及目的蛋白与其他分子的相互作用等后续研究。传统的酶切连接方法不能满足大规模克隆的需要,只有通过一种高效的对载体和宿主没有依赖性的克隆系统,才能完成这项繁杂的工作。Gateway 基因大规模克隆法利用 λ 噬菌体进行位点特异性 DNA 片段重组,实现了不需要传统的酶切连接过程的基因快速克隆和载体间平行转移,为大规模克隆基因组注释的可译框架提供了保障。

Gateway 克隆技术主要包括 TOPO 反应和 LR 反应。TOPO 反应将目的基因 PCR 产物连入 Entry 载体,该载体上的 CCCTT 被拓扑异构酶所识别,切开后通过 274 位酪氨酸与切口处的磷酸基团形成共价键, 从而将该酶偶联在载体上。加入 PCR 产物后,形成的 3′突出端 GTGG,攻击 PCR 产物的互补性末端并与接头序列 CACC 退火,切去 GTGG 后使 PCR 产物以正确方向连入 Entry 载体(图 5-24a)。

LR 反应被用于将目的片段从 Entry 载体中重组入表达载体。Entry 载体上基因两端具有 attL1 和 attL2 位点,目的载体上含有 attR1 和 attR2 位点,在重组蛋白的作用下发生定向重组,形成新的位点 attB1 和 attB2,将目的基因转移到表达载体中(图 5-24b)。

### 5.5.4 基因的图位克隆法

基因的图位克隆法(map-based cloning)是分离未知性状目的基因的一种好方法。从理论上说,所有具有某种表现型的基因都可以通过该方法克隆得到。首先,通过构建遗传连锁图,将目的基因定位到某个染色体的特定位点,并在其两侧确定紧密连锁的 RFLP 或 RAPD 分子标记。其次,通过对许多不同的生态型及大量限制性内切酶和杂交探针的分析,找出与目的基因距离最近的 RFLP 标记, 通过染色体步移技术将位于这两个标记之间的基因片段克隆并分离出来(图 5-25)。然后,根据基因功能互作原理鉴定目的基因。在 RFLP 作图中,连锁

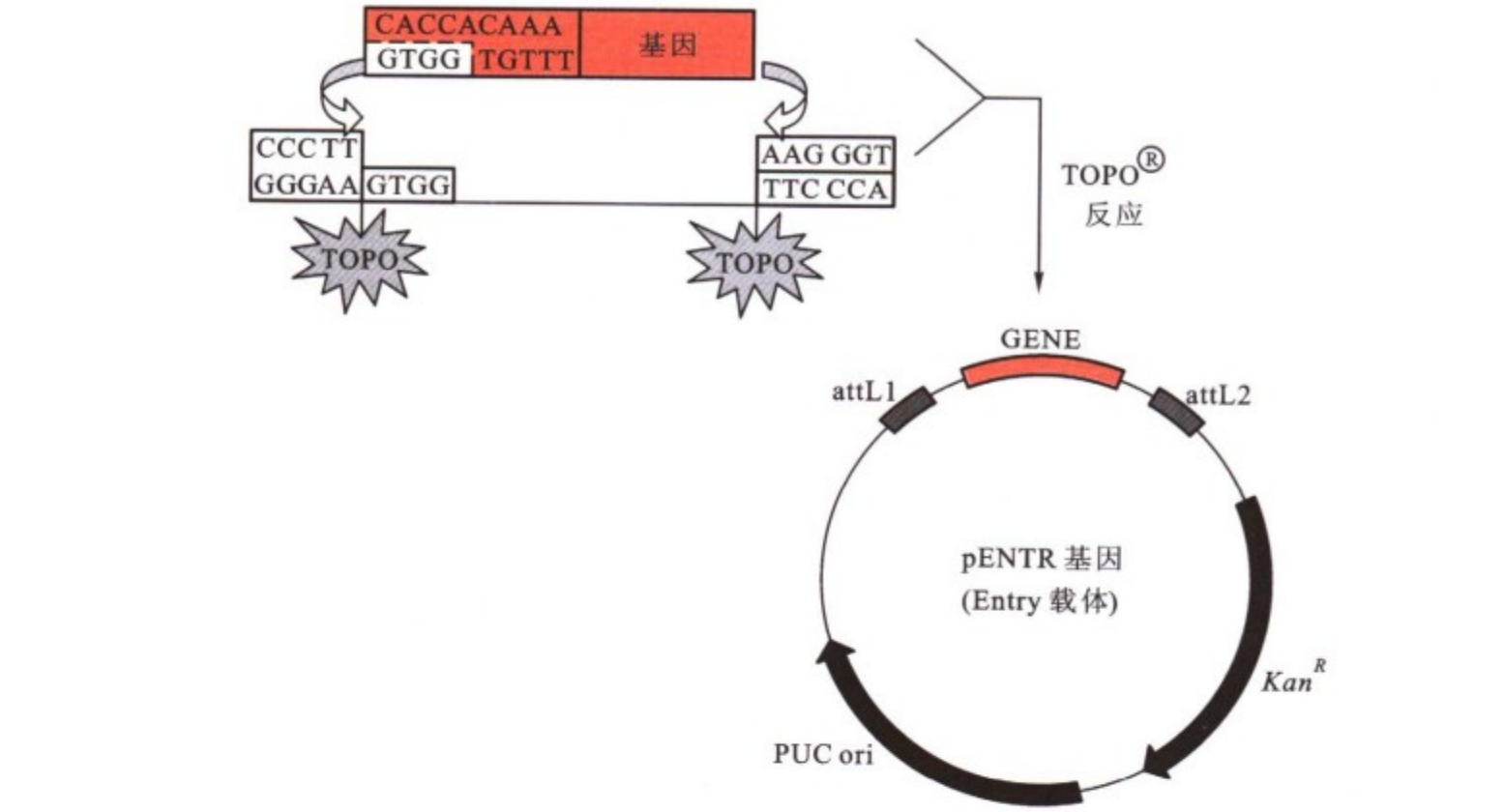



**图 5-24 Gateway 大规模克隆策略** (a)TOPO 反应,将由 PCR 得到的目的基因连入 Entry 载体;(b)LR 反应,通过位点特异性重组将目的基因从 Entry 载体连入表达载体。

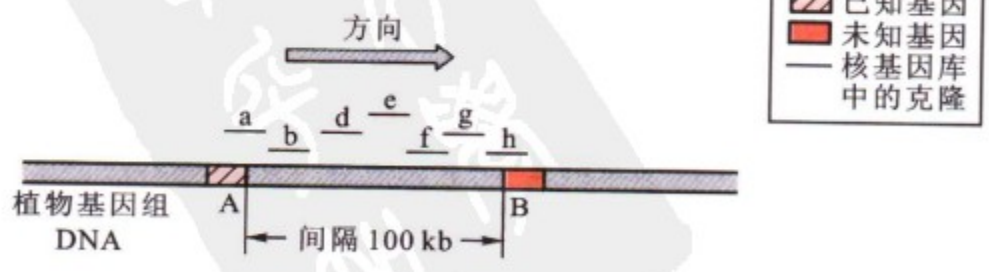



**图 5-25 染色体步移法克隆基因示意图**

距离是根据重组率来计算的，1 cM(厘摩)相当于1%的重组率。人类基因组中，1 cM≈1 000 kb；拟南芥中，1 cM≈290 kb；小麦中，1 cM≈3 500 kb。

由于水稻基因组比较小，与其他单子叶植物有较高的共线性现象，而且基因组序列分析工作已经基本完成，基因转化系统也比较完善，所以成为单子叶植物功能基因研究的模式系统。近年来，我国科学家利用自主筛选的水稻突变体和高密度分子标记遗传图谱，应用该技术成功地分离出多个具有重要生理功能的水稻基因，包括控制水稻分蘖的关键基因 *MOC1* 和控制植株茎秆强度的脆杆基因 *BC1*。图 5-26 介绍了图位法克隆 *BC1* 基因的实验流程。

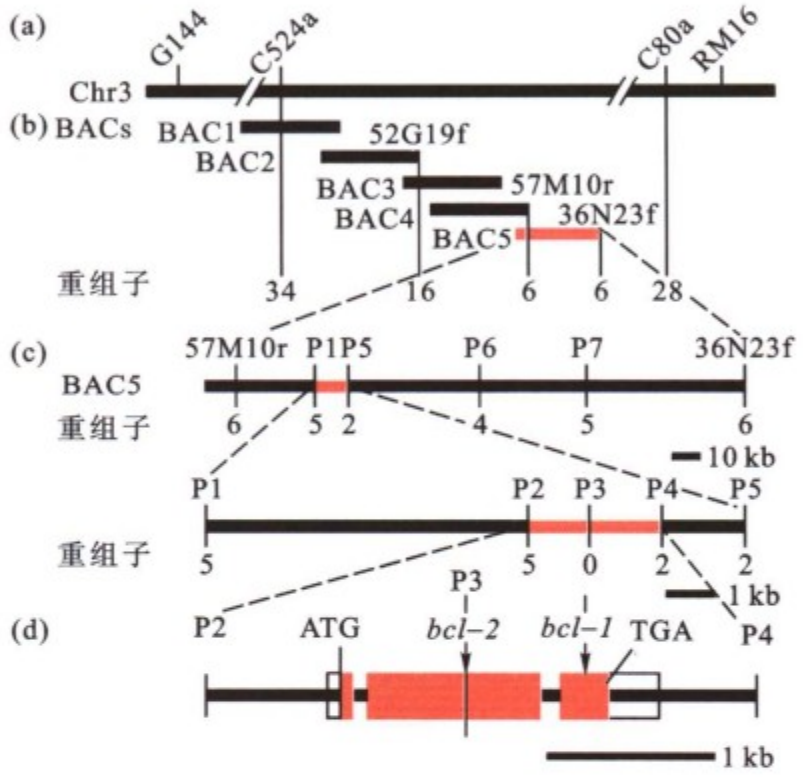



**图 5-26 用图位克隆法获得水稻脆杆基因 *BC1*** (a)首先将 *BC1* 位点定位于水稻3号染色体(Chr3)着丝粒区的分子标记 C524a 和 RM16 之间；(b)覆盖 *BC1* 位点的 BAC 大片段，数字表示从 7 068 个 *bcl*-2 突变体 F2 植株中鉴定出的重组子数目；(c)*BC1* 位点的精细定位，*BC1* 位点被定位于 CAPS(cleaved amplified polymorphic sequence) 分子标记 P2 和 P4 之间 3.3 kb 的 DNA 区段，并与 P3 标记共分离；(d)*BC1* 基因结构，黑色矩框为编码区，白色矩框为 5′和 3′非翻译区，矩框之间的线段为内含子。*Bcl*-1 和 *bcl*-2 表示两个脆杆突变体的突变位点。

## 5.6 蛋白质组与蛋白质组学技术

1995 年，澳大利亚科学家首次在电泳杂志(Electrophoresis)上发表关于蛋白质组学 (proteome) 概念的论文。蛋白质组学是蛋白质 (protein) 和基因组(genome)研究在形式和内容两方面的完美组合，该技术致力于研究某一物种、个体、器官、组织或细胞在特定条件、特定时间所表达的全部蛋白质图谱。蛋白质组与基因组既相互对应又存在显著不同，因为基因组是确定的，组成某个体

的所有细胞共同享有固定的基因组,而各个基因的表达调控及表达程度却会根据时间、空间和环境条件发生显著的变化,所以,不同器官、组织或细胞内拥有不同的蛋白质组。由于蛋白质分离(改进后的双向电泳技术和高效液相色谱技术)和鉴定技术(现代质谱)的快速更新,由于基因组学研究和生物信息学的交叉渗透,蛋白质组学研究在近年来获得了长足的进展。

### 5.6.1 双向电泳技术

到目前为止,虽然科学家已经发明了多种蛋白质分离技术,双向电泳技术仍然是分离大量混合蛋白质组分的最有效方法。科学家大约在 1975 年首次建立了等电聚焦(isoelectric focusing,IEF)及 SDS-聚丙烯酰胺凝胶(SDS-PAGE)双向电泳技术(two-dimensional electrophoresis,2-D)(图 5-27)。该方法依赖于蛋白质的两个重要特性,等电点和相对分子质量。蛋白质是两性分子,在不同 pH 的缓冲液中表现出不同的带电性,因此,在电流的作用下,在以两性电解质为介质的电泳体系中,不同等电点的蛋白质会聚集在介质上不同的区域(等电点)而被分离。蛋白质的相对分子质量决定了 SDS-蛋白复合物在凝胶电泳中的迁移率。聚丙烯酰胺凝胶中的去垢剂 SDS 带有大量的负电荷,与之相比,蛋白质所带电荷量可以忽略不计。因此,蛋白质在 SDS 凝胶电场中的运动速度(速度与相对分子质量成反比) 和距离完全取决于其相对分子质量而不受其所带电荷的影响,不同相对分子质量的蛋白质将位于凝胶的不同区段而得到分离。20 世纪 80 年代发明的固相化 pH 梯度技术(immobilized pH gradients,IPG)克服了载体两性电解质向阴极漂移等许多缺点,逐步建立了稳定的可精确设定的 pH 梯度,增大了蛋白上样量,提高了双向凝胶电泳结果的可重复性。同时随着各种试剂质量的不断提高和新试剂的开发与应用,双向凝胶的分辨率空前提高,其最大分辨率已达到每块胶 10 000 个蛋白点。

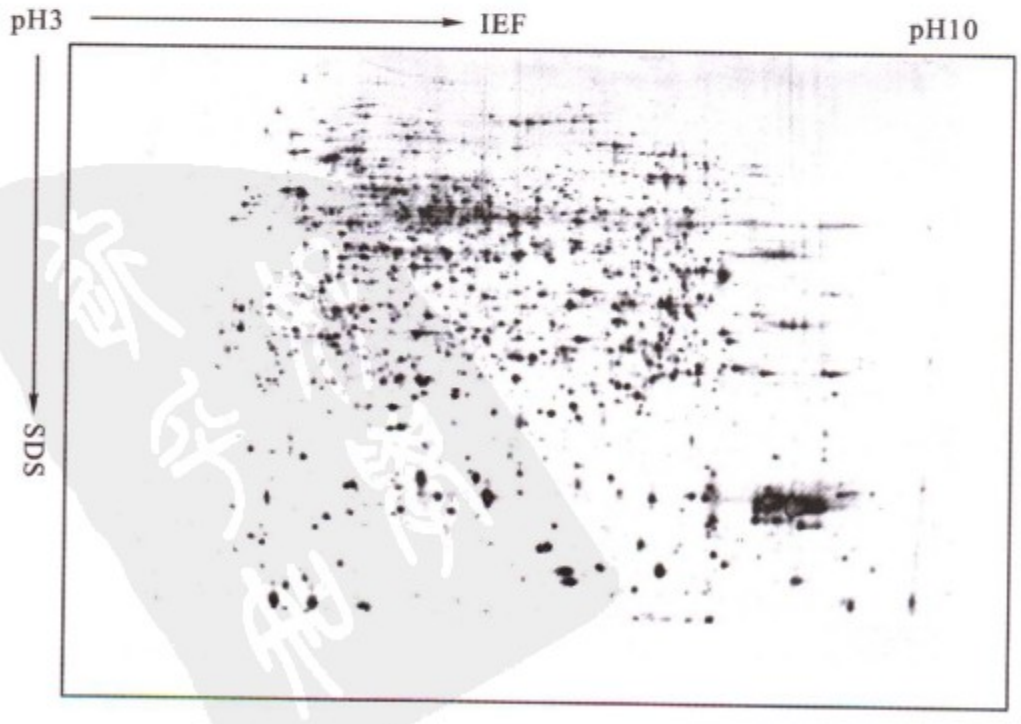



图 5-27 用 2-D 胶展示拟南芥种子萌发 48h 后的蛋白质

## 5.6.2 蛋白质印迹法

分子生物学中最常用的蛋白质技术可能是蛋白质印迹法（Western blotting），又称免疫印迹法（immunoblotling），该技术是20世纪70年代末80年代初在蛋白质凝胶电泳和固相免疫测定基础上发展起来的蛋白质检测技术，为检测样品中是否存在蛋白抗原提供了一种可靠的方法。该法鉴定蛋白质的原理是根据被测蛋白能与特异性抗体结合的特性和该蛋白的相对分子质量。免疫印迹法亦可用于测定抗原的其他特性，如抗原的相对多度或与其他已知抗原的关系。它也是评价新抗体特异性的一种有效方法。该法具有高效、简便、灵敏等特点。

免疫印迹法程序可分为五个步骤：①蛋白样品的制备；②经过SDS-PAGE分离样品；③分离的蛋白转移到膜载体上，转移后首先将膜上未反应的位点封闭起来以抑制抗体的非特异性吸附；④用固定在膜上的蛋白质作为抗原，与对应的非标记抗体（一抗）结合；⑤洗去未结合的一抗，加入酶偶联或放射性同位素标记的二抗，通过显色或放射自显影法检测凝胶中的蛋白成分（图5-28）。

影响免疫印迹法成败的一个主要因素是抗原分子中可被抗体识别的表位性质。所以在选择抗体（一抗）时要考虑所选抗体是否能识别凝胶电泳后转印到膜上的变性蛋白，还有所选抗体是否会引起交叉反应（表5-5）。第二个影响因素是蛋白原液中抗原的浓度。对于中等相对分子质量的蛋白质（50 000左右）其浓度需大于0.1 ng才能被检出。如果要检测更低量的蛋白质，样品要做进一步的纯化。

用于检测的二级抗体一般分为以下三种：

1. 放射性标记的抗体

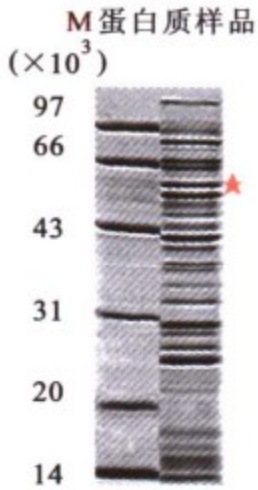
最早的Western blotting都采用这种方法，这也是迄今为止定量最准确的方法。由于使用放射性同位素，有一定的危险性，现在已逐渐被非放射性检测系统所取代。

2. 与酶偶联的抗体

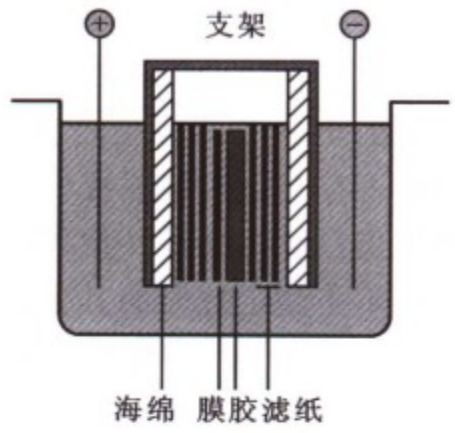
主要包括辣根过氧化物酶和碱性磷酸酯酶，如果选用辣根过氧化物酶标记的二抗，就可以用化学发光法检测。在过氧化氢存在的条件下，辣根过氧化物酶催化鲁米诺氧化即可发光，在暗室用X线片曝光法可以检测到信号（图5-29）。如果选用碱性磷酸酶标记的二抗和溴氯吲哚磷酸盐/硝基氮蓝四唑（BCIP/NBT）底物，就可以在酶结合部位产生肉眼看得到的黑紫色沉淀。BCIP/NBT底物能特征性地产生清晰条带，因此背景很干净。反应可以被EDTA终止。

3. 与生物素相偶联的抗体

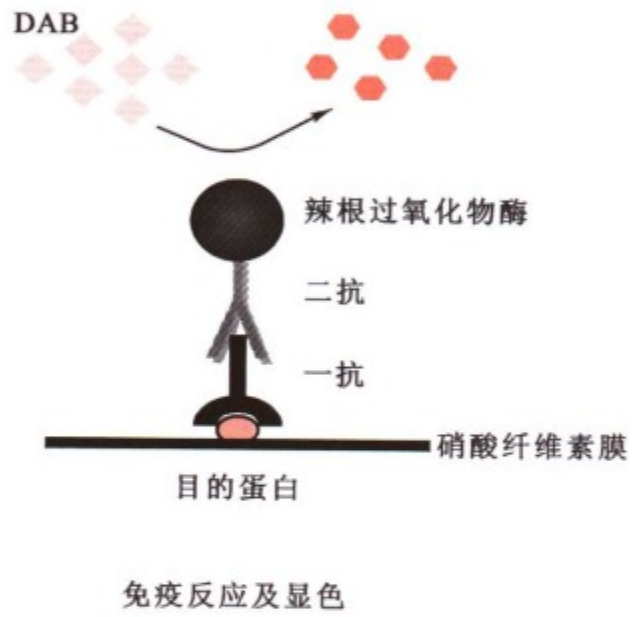
生物素是可溶性的维生素，能与抗生物素蛋白高亲和力结合，后者再与报道酶相结合，以确定抗生物-生物素-靶蛋白复合物的位置并进行定量分析。



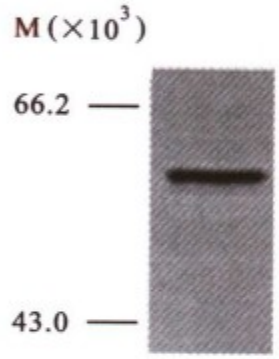

蛋白电泳(打★号者为目的蛋白)




转膜




免疫反应及显色




目的蛋白 western blotting 结果

图 5-28 免疫印迹法示意图

表 5-5 抗体的选择

| | 多克隆抗体 | 单克隆抗体 | 混合单克隆抗体 |
|---|---|---|---|
| 信号强度 | 较好 | 视不同抗体而异 | 最佳 |
| 特异性 | 良好,但有一定的背景 | 最佳,但有交叉反应 | 最佳 |
| 优点 | 多数能识别变性抗原 | 特异性好,抗体来源不受限制 | 信号强,特异性好,抗体来源不受限制 |
| 缺点 | 不易重复,有时背景较深 | 多数不能识别变性抗原 | 不易获得 |

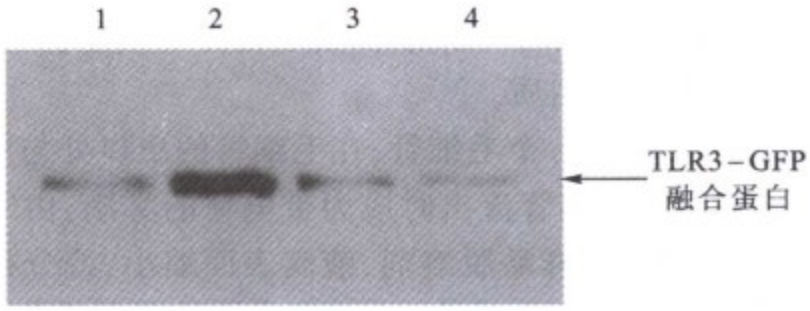



图 5-29 用辣根过氧化物酶显色法和抗 GFP 的抗体检测外源 TLR3-GFP融合蛋白在同一细胞系不同克隆株中的表达水平(1,2,3,4 表示不同的克隆株)

## 5.6.3 蛋白质的质谱分析技术

蛋白质组学中最有意义的突破是用生物质谱鉴定电泳分离后的蛋白质。该技术是一种超灵敏的鉴定技术,从理论上说,仅需极少量的样品就可以获得错误率低于百万分之一的精确结果。可将现行的质谱仪简单地分为三个连续的组成部分,即离子源,离子分离区和检测器。目前比较常用的是基质辅助的激光解析电离-飞行时间质谱 (matrix-assisted laser desorption ionization time-of-flight spectrometry,MALDI-TOF) 和电喷雾质谱 (electrospray ionisation,ESI-MS)。MALDI-TOF 的工作原理是将从 2-D 胶中分离得到的或其他来源的蛋白质酶解成小肽段后与基质(主要是有机酸)混合,将样品混合物点到金属靶表面上并使之干燥结晶,然后用激光轰击,将呈离子化气体状态的待分析物从靶表面喷射出去。离子化的气体中每个分子带有一个或更多的正电荷,这些气体肽段在电场中被加速后到达检测器的时间由肽段的质量和其所带电荷数的比值($m/z$)决定(图 5-30)。通过分析那些只带有单电荷的气体化肽段便能掌握该肽段的精确的分子质量,再根据不同质核比离子化肽段到达检测器的时间绘出质谱图。

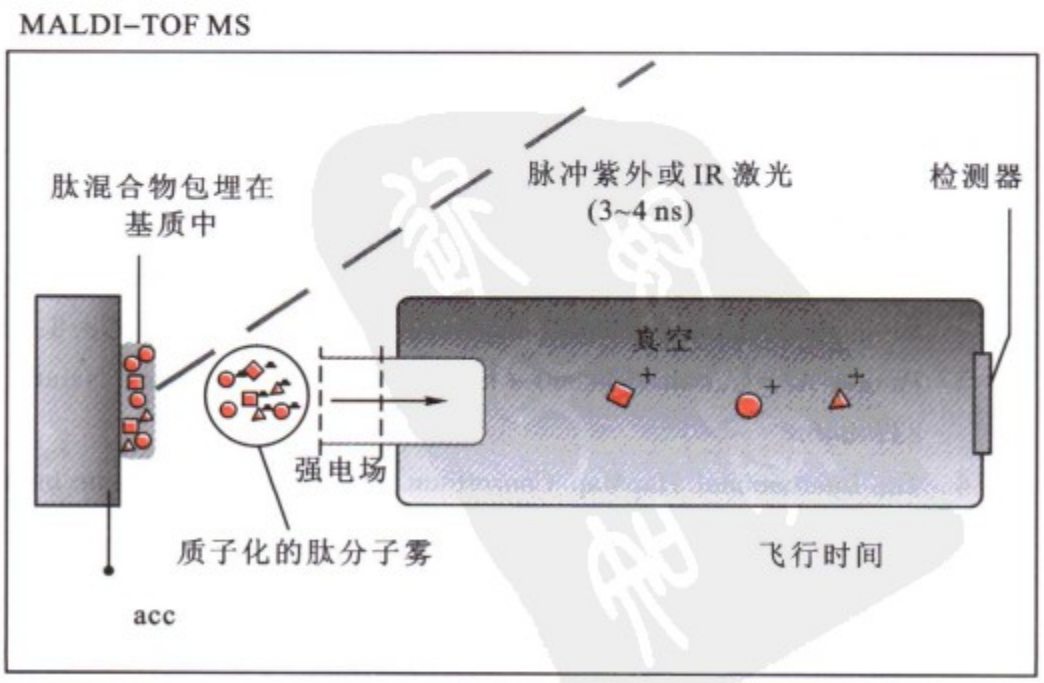



图 5-30 MALDI-TOF 分子质谱仪原理图

比较肽质量指纹图谱与数据库中不同肽段的理论分子质量,就能找到与肽指纹图谱有不同程度相似性的蛋白质。

典型的电喷雾质谱源是一个毛细管,在毛细管的出口处与反电极之间施加较高的电压,样品溶液从毛细管流出时成为带电荷的雾滴,由于电喷物离子源处有加热的辅助气体($N_2$)使样品脱溶剂,液滴表面缩小,待分析的蛋白质逐渐以单电荷或多电荷形式成为气态物质。电喷雾电离使蛋白质分子比较容易成为多电荷离子,将大多数蛋白的质荷比降低到质谱仪的检测范畴之内,提高了质谱仪的分析能力。ESI-QqTOF是在ESI-TOF和三级四级杆质谱仪基础上发展起来的,ESI-QqTOF中的ESI是指电喷物离子源,Q代表四级杆质量分析器,q代表碰撞室,主要起离子碰撞活化和聚焦作用。TOF代表时间质量分析器。该仪器充分结合了四级杆和TOF质量分析器的优点,具有分辨率高和灵敏度高的特点,特别适合对未知肽的检测。

## 思考题

1. 哪些重要的科学发现和实验推动了DNA重组技术的产生和发展?
2. 如何理解PCR扩增的原理和过程?
3. 简述定量PCR的原理和过程。
4. 基因组DNA文库和cDNA文库在构建原理和用途上的主要区别是什么?
5. 基因克隆的方法主要有哪几种? 简述各种方法的原理和用途。
6. 在基因操作实践中有哪些检测核酸和蛋白质相对分子质量的方法?
7. 蛋白质组学的研究对象和目的是什么? 主要有哪些技术和方法?
8. SNP作为第三代遗传标记的优点是什么?
9. 基因分型的方法有哪些? 简述其原理。
10. 已知一个cDNA 3′端的部分序列,请设计实验流程得到该基因的全长cDNA。

## 主要参考文献


1. Clark D. Molecular Biology. Elsevier Academic Press, 2005.
2. Landegren U, Nilsson M, Kwok P Y. Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. Genome Res., 1998, 8: 769–776.
3. Pandey A, Mann M. Proteomics to study genes and genomes. Nature, 2000, 405: 837–846.
4. The International HapMap Consortium. A haplotype map of the human genome. Nature, 2005, 437: 1299–1320.

# 6 分子生物学研究法(下)
## ——基因功能研究技术


- 6.1 基因表达研究技术
- 6.2 基因敲除技术
- 6.3 蛋白质及 RNA 相互作用技术
- 6.4 基因芯片及数据分析
- 6.5 利用酵母鉴定靶基因功能
- 6.6 其他分子生物学技术


随着酵母、拟南芥、水稻、小鼠、人类等基因组序列相继地被测定完成,我们对生物体的认识又前进了一大步,但是大量的序列信息也给我们提出了新的挑战。如何利用这些序列信息,通过生物化学、分子生物学等方法研究基因的功能,从而进一步了解生物体内各种生理过程,研究生物体生长发育的调节机制,以及各种疾病的发生、发展过程是新时期生物学家所面临的主要问题。基因表达系列分析技术、原位杂交技术、基因芯片技术为研究单个基因或多个基因在生物体的某些特定发育阶段或在不同环境条件下的表达模式提供了强有力的手段。用基因定点突变(site-directed mutagenesis)技术、基因敲除技术、RNAi 技术可以全部或部分抑制基因的表达,通过观察靶基因缺失后生物体的表型变化研究基因功能。酵母单杂交、双杂交技术,四分体技术等也是研究蛋白质相互作用、蛋白质-DNA 相互作用等的重要手段。随着分子生物学技术的发展,研究者可以在活细胞内和细胞外研究蛋白质之间的相互作用,为认识信号转导通路、蛋白质翻译后修饰加工等提供了强有力的技术支持。本章将主要介绍研究基因功能的各种分子生物学技术和方法。

## 6.1 基因表达研究技术

### 6.1.1 基因表达系列分析技术

基因表达系列分析技术 (serial analysis of gene expression, SAGE) 是一种以 DNA 序列测定为基础定量分析全基因组表达模

式的技术,能够直接读出任何一种细胞类型或组织的基因表达信息。在转录组水平上,任何长度超过9~10个碱基的核苷酸片段都可能代表一种特异性的转录产物。因此,用特定限制性核酸内切酶分离转录产物中具有基因特异性的9~10个碱基的核苷酸序列并制成标签。将这些序列标签连接、克隆、测序后,根据其占总标签数的比例即可分析其对应编码基因的表达频率。

SAGE的操作过程如图6-1所示,主要包括:分离提取mRNA,反转录反应合成cDNA双链,用特定限制性内切酶—锚定酶(anchoring enzyme,AE)如

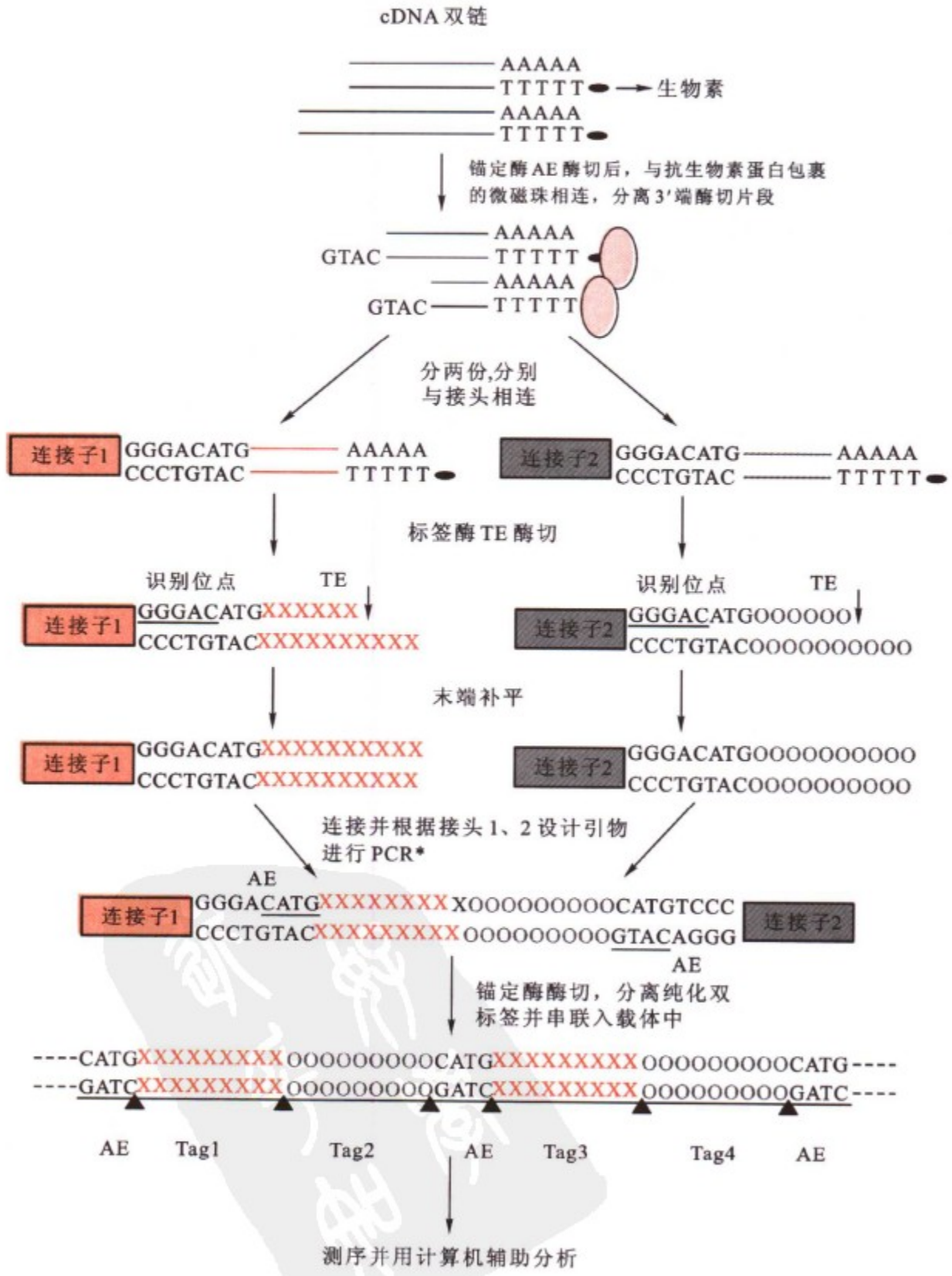



图6-1 常规SAGE方法(short SAGE)基本流程 *连接子1和连接子2的5′端分别加有氨基,以防止5′端相连接。

*Nla*Ⅲ进行消化。锚定酶识别位点多为 4 个碱基,以保证各个转录产物中至少含有一个酶切位点。分离 3′端酶解片段并将其分为两份,分别与连接子 1 和连接子 2[均含有ⅡS 类标签酶(tagging enzyme,TE)的识别位点]结合,用标签酶(一种Ⅱ类限制酶,它能在距离识别位点数个或十数个碱基的位置切割 DNA 双链)如 *Bsm*F1 酶切,将含有连接子 1 和 2 的样品混合,形成双标签。再根据连接子 1 和 2 的序列设计引物,进行 PCR 扩增,经 PAGE 电泳回收目的条带,用 AE 酶切、回收、连接双标签并进行克隆、筛选和测序,进行基因的鉴定和表达频率分析。选择不同的 TE 可以改变标签的长度。设计连接子时要使 3′端与 AE 酶切的末端能够互补,以保证 1 和 2 以首尾相接进行克隆。一般情况下,每个克隆含有的标签在 10~50。

随着越来越多的生物基因组序列被测定完成,人们发现传统的 SAGE 技术用 14 个碱基(10 个基因特异碱基序列+4 个碱基的酶切位点)的标签代表一个基因并不能覆盖像人类这样的大基因组内所有的基因序列,因而又发展了 LongSAGE 技术,将相对应的传统 SAGE 方法称为 ShortSAGE。LongSAGE 标签来自转录物 3′端一段 21 bp 的序列,可以进行快速分析并与基因组序列数据相匹配。LongSAGE 方法与 ShortSAGE 方法类似,只是用了不同的ⅡS 类标签酶(*Mme* Ⅰ),并将程序做了相应修改。

## 6.1.2 RNA 的选择性剪接技术

RNA 的选择性剪接是指用不同的剪接方式(选择不同的剪接位点组合)从一个 mRNA 前体产生不同的 mRNA 剪接异构体的过程。一般将选择性剪接分为如下几类:平衡剪接、5′选择性剪切、3′选择性剪接、外显子遗漏性剪接及相互排斥性剪接(图 6-2)。一般用 RT-PCR 的方法研究一个基因是否存在选择性剪

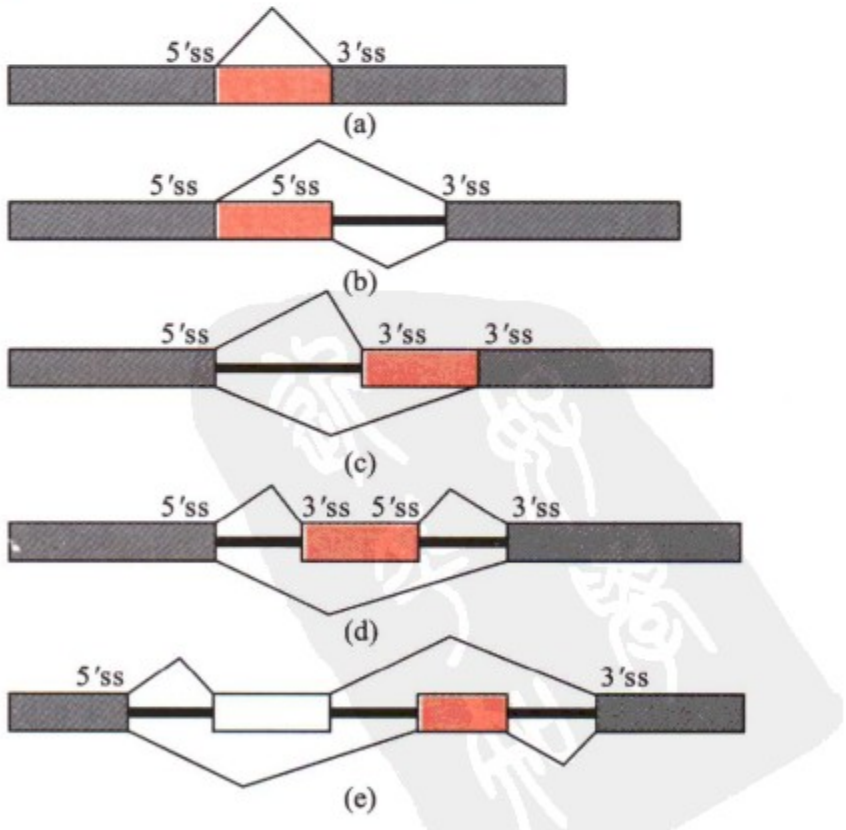



**图 6-2 选择性剪接的不同类型** (a)平衡剪接;(b)5′选择性剪接;(c)3′选择性剪接;(d)外显子遗漏;(e)相互排斥性剪接。

接。首先以 cDNA 两端特异引物或来自不同外显子的引物序列在不同组织来源的 RNA 样品中进行扩增,观察 PCR 产物大小是否存在差异。如果发现差异,测序后可看出这种差异是否来自于选择性剪接。图 6–3 为拟南芥中发现的有选择性剪接的 5 个转录调控因子基因的物理图谱。选择性剪接使一个基因翻译为多种蛋白质序列,是基因表达多样性的重要表现形式。分析人类基因组数据发现,有 60%的基因在表达过程中可通过选择性剪接产生各种形式的 mRNA。最新的研究表明,果蝇的 *Dscam* 基因最多能够产生 38 016 种不同形式的剪接体(图 6–4)。

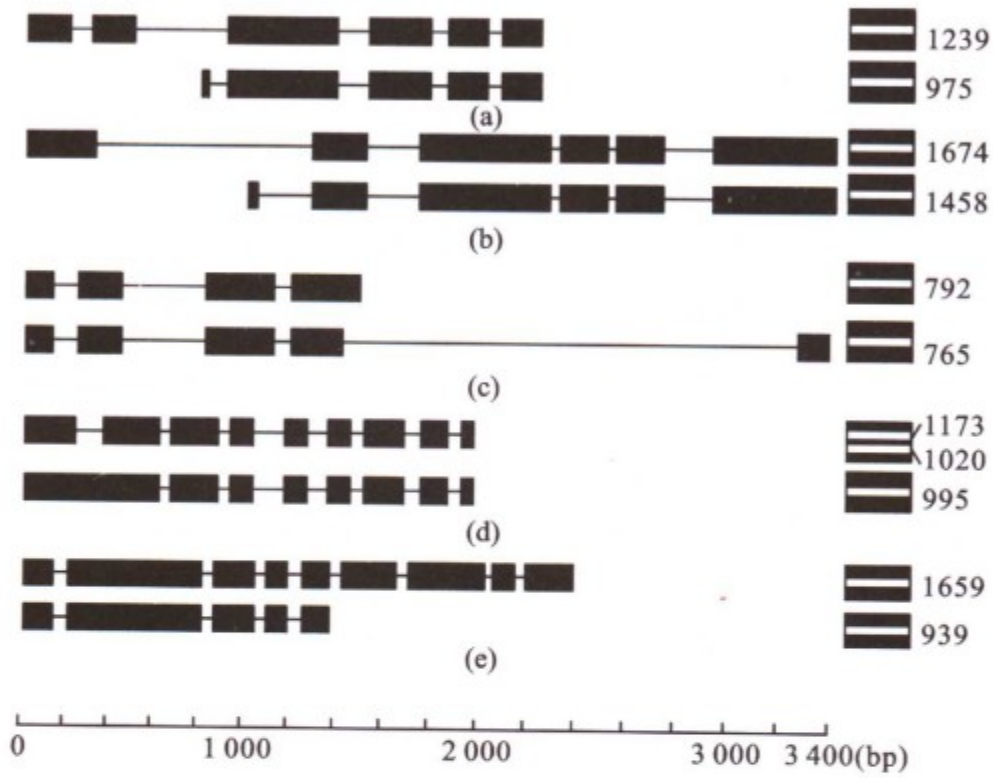



**图 6–3 RT–PCR 检测 5 个拟南芥转录调控因子基因的选择性剪切**

(a),(b),(c),(d) 和 (e) 分别代表 Atlg72050,At4g26640,Atlg26260,At4g29930 和 At3g46590。

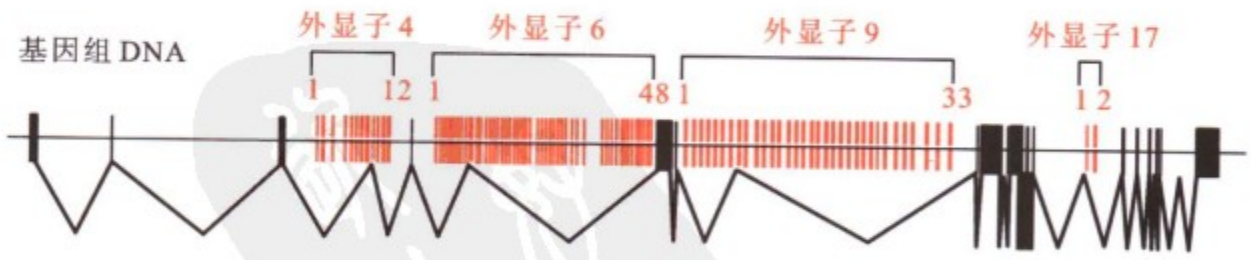


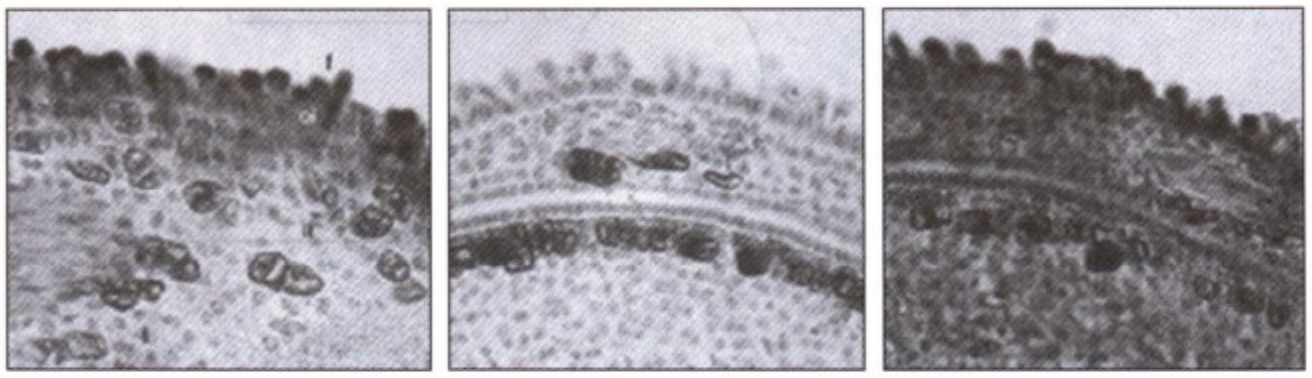

**图 6–4 果蝇(*Drosophila melanogaster*)的 *Dscam* 基因可能通过可变剪接产生 38 000 多种可能的 mRNA 异构体**

*Dscam* 基因编码一个神经无轴突定向受体,它在细胞外有一个由 10 个免疫球蛋白重复序列组成的结构域,第 2,3,7 个免疫球蛋白重复序列分别由第 4,6,9 号外显子编码,4 号外显子有 12 个变异体,6 号外显子有 48 个变异体,9 号外显子有 33 个变异体,再加上 17 号外显子的 2 个变异体。每个成熟的 Dscam mRNA 分别只有一个由 4,6,9,17 号外显子组成的变异体,由此理论推测 *Dscam* 基因共有 12×48×33×2=38 016 个可能的剪接异构体。

由于选择性剪切与细胞生理、发育调节以及肿瘤的发生、转移等有密切关系，阐明基因的选择性剪切机制是了解动植物个体发育和基因功能的重要环节。因此，发现新的可变剪切异构体，确定每个异构体的独特功能和生物学意义并阐明其调节机制，是功能基因组时代的一个重要特征。

### 6.1.3 原位杂交技术

原位杂交(in situ hybridization，ISH) 是用标记的核酸探针，经放射自显影或非放射检测体系，在组织、细胞、间期核及染色体上对核酸进行定位和相对定量研究的一种手段，通常分为 RNA 原位杂交和染色体原位杂交两大类。RNA 原位杂交用放射性或非放射性(如地高辛、生物素等)标记的特异性探针与被固定的组织切片反应，若细胞中存在与探针互补的 mRNA 分子，两者杂交产生双链 RNA，就可通过检测放射性标记或酶促免疫反应显色，对该基因的表达产物在细胞水平上做出定性定量分析(图 6-5)。



(a) (b) (c)

**图 6-5 用组织原位杂交技术检测棉花纤维特异基因的表达** (a)mRNA 的互补链，杂交信号集中于纤维细胞中；(b)负对照，mRNA 的同义链无杂交信号；(c)正对照，UBQ10 杂交信号遍布于整个胚珠。

荧光原位杂交(fluorescence in situ hybridization，FISH)技术首先对寡核苷酸探针做特殊的修饰和标记，然后用原位杂交法与靶染色体或 DNA 上特定的序列结合，再通过与荧光素分子相偶联的单克隆抗体来确定该 DNA 序列在染色体上的位置。FISH 技术不需要放射性同位素，实验周期短，检测灵敏度高，若用经过不同修饰的核苷酸分子标记不同的 DNA 探针，还可能在同一张切片上观察几种 DNA 探针的定位，得到相应位置和排列顺序的综合信息。

### 6.1.4 基因定点突变技术

定点突变(site-directed mutagenesis)是重组 DNA 进化的基础，该方法通过改变基因特定位点核苷酸序列来改变所编码的氨基酸序列，常用于研究某个(些)氨基酸残基对蛋白质的结构、催化活性以及结合配体能力的影响，也可用于改造 DNA 调控元件特征序列，修饰表达载体，引入新的酶切位点等。由于迄今尚未建立用于精确预测特定氨基酸残基变化对整个蛋白结构及活

性影响的模型,使选择氨基酸突变的位点存在一定的盲目性。因为即使知道了一个蛋白质的三维结构,人们仍然无法了解某个氨基酸残基对其高级结构的影响。少数几个氨基酸残基的改变,有时根本不影响靶蛋白的功能,有时影响了靶蛋白的活性位点,有时还能从根本上改变整个蛋白的高级结构。所以,基因的定点突变为我们进一步了解蛋白质的结构与功能的关系提供了重要的手段。

最早在20世纪70年代初,进行小噬菌体ΦX174单链基因组易突变位点图谱分析工作时就认识到了基因定点突变的可能性。在人工成功合成寡聚核苷酸、获得高品质DNA聚合酶和DNA连接酶的基础上,Smith及其同事建立了体外寡核苷酸介导的DNA突变技术(图6-6)。PCR技术的出现大大促进了定点突变技术的发展,简化了实验操作程序,提高了突变效率。科学家主要采用两种PCR方法,重叠延伸技术和大引物诱变法,在基因序列中进行定点突变。

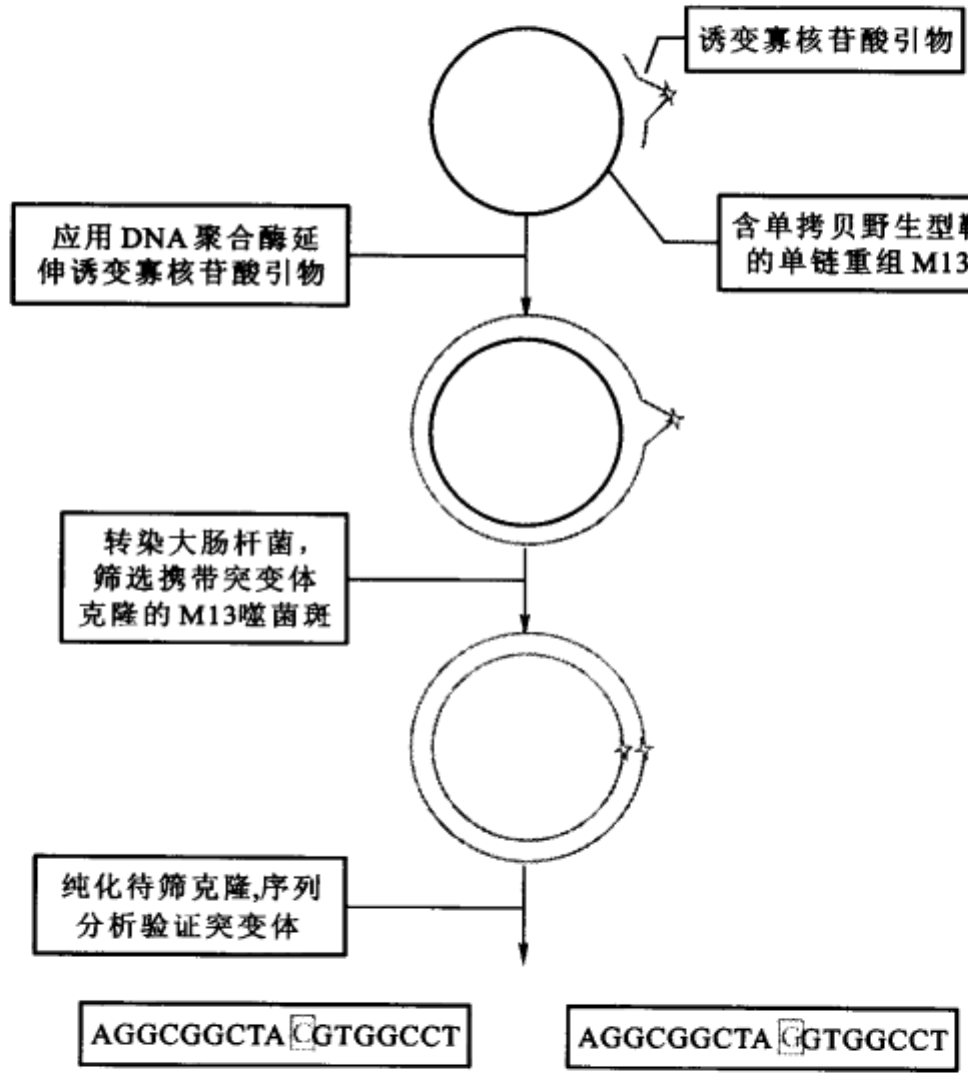



图6-6 寡核苷酸介导的DNA突变技术

图6-7阐述了重叠延伸介导的定点诱变机制。首先将模板DNA分别与引物对1(正向诱变引物FM和反向引物R2)和2(正向引物F2和反向诱变引物RM)退火,通过PCR1和2反应扩增出两种靶基因片段。FMR2和RMF2片段在重叠区发生退火,用DNA聚合酶补平缺口,形成全长双链DNA,进行PCR3扩增。最后,用引物F2和R2扩增出带有突变位点的全长DNA片段(PCR4)。

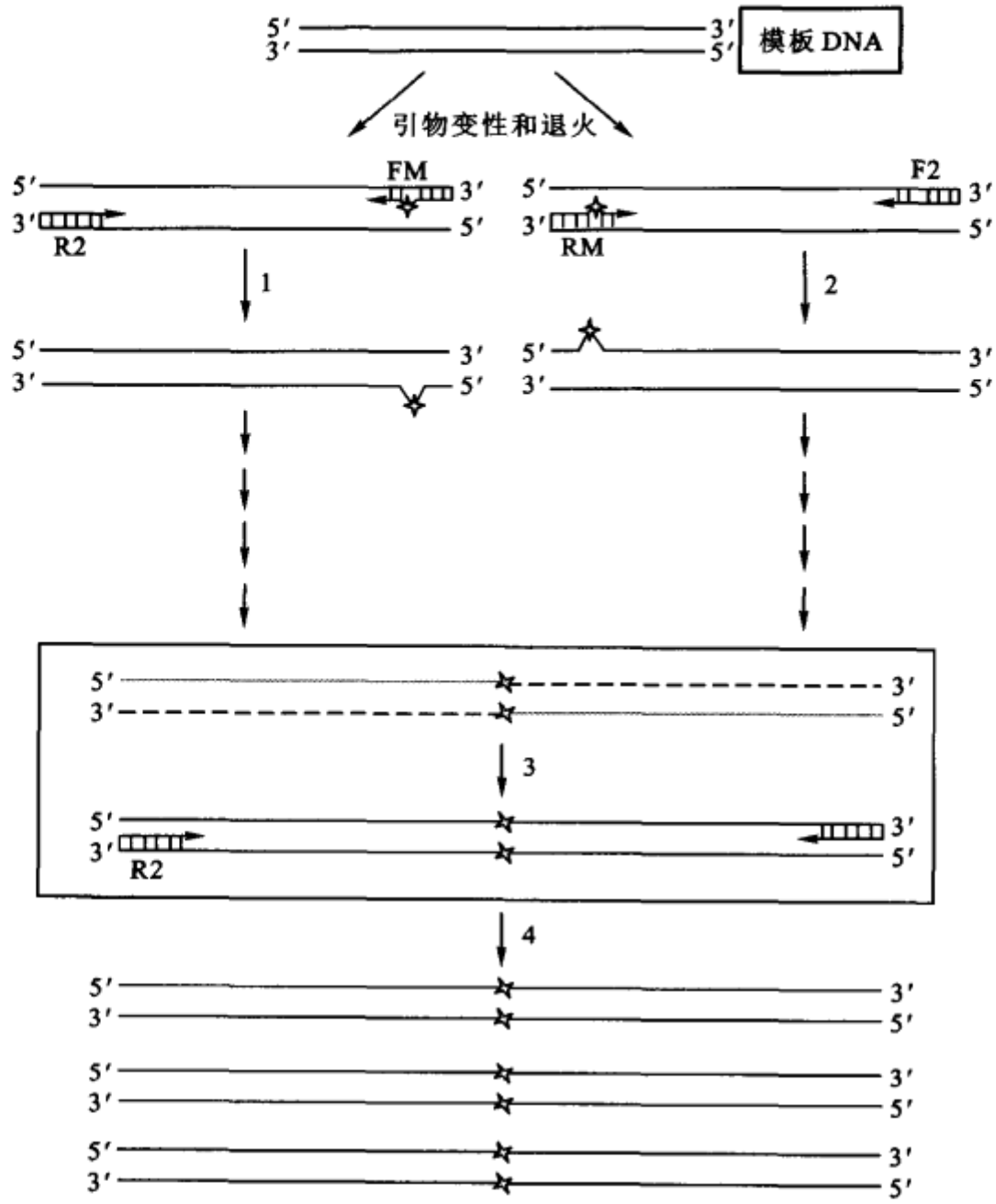



**图 6-7 重叠延伸介导的定点诱变** 模板 DNA 分别与引物对 1(正向诱变引物 FM 和反向引物 R2)及引物对 2(正向引物 F2 和反向诱变引物 RM)退火,通过 PCR1 和 2 反应扩增出两种靶基因片段。在 PCR3 反应中变性后的 FMR2 和 RMF2 片段在重叠区发生退火,按虚线所示进行延伸,并形成全长双链 DNA。PCR4 使用引物 F2 和 R2 扩增出全长突变 DNA。

大引物诱变法首先用正向突变引物 (M) 和反向引物 (R1), 扩增模板 DNA 产生双链大引物(PCR1),与野生型 DNA 分子混合后退火并使之复性,第二轮 PCR 中加入正向引物(F2),与 PCR1 中产生的一条互补链配对,扩增产生带有突变的双链 DNA (图 6-8)。由于 F2 的退火温度显著高于第一轮 PCR 所使用的引物 M 和 R1,因此,可以忽略引物 M 和 R1 在本轮反应中所造成的干扰。

与经典定点突变方法相比,PCR 介导的定点突变方法具有明显的优势:①突变体回收率高,以至于有时不需要进行突变体筛选;②能用双链 DNA 作为模板,可以在任何位点引入突变;③可在同一试管中完成所有反应;④快速简便, 无需在噬菌体 M13 载体上进行分子克隆。所以,PCR 介导的定点突变方法

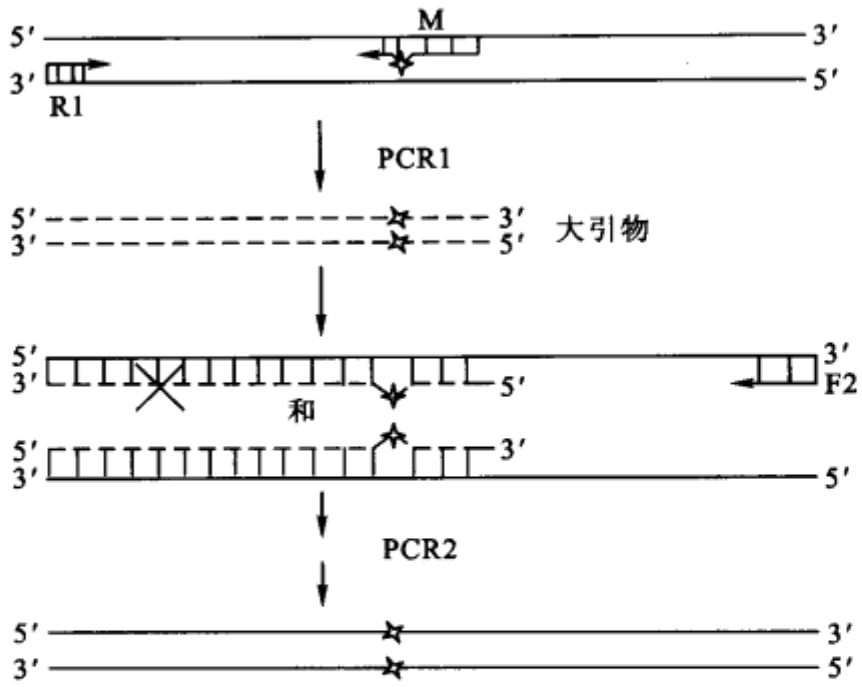



图 6-8 大引物诱变法示意图

正在成为定点突变的主流。除 DNA 序列分析之外,也可用变性梯度凝胶电泳和单链构象多态性等方法检测产生突变的 DNA 片段。

## 6.2 基因敲除技术

### 6.2.1 基本原理

经典遗传学(forward genetics)从一个突变体的表型出发,研究其基因型,进而找出该基因的编码序列。在后基因组时代,大规模基因功能的研究正成为生命科学研究的热点,现代遗传学(reverse genetics,反向遗传学)首先从基因序列出发,推测其表现型,进而推导出该基因的功能。基因敲除(gene knock-out)又称基因打靶,该技术通过外源 DNA 与染色体 DNA 之间的同源重组,进行精确的定点修饰和基因改造,具有专一性强、染色体 DNA 可与目的片段共同稳定遗传等特点。

1985 年,科学家利用同源重组将一段外源质粒 pΔβ117 插入到人类染色体 DNA β-珠蛋白位点,首次在哺乳动物细胞中进行基因打靶并获得成功。目前,在胚胎干细胞中进行同源重组已经成为遗传修饰小鼠染色体组基因位点的常规技术。通过对这些基因敲除小鼠的表形分析,新发现了许多与人类疾病相关的基因。

基因敲除技术分为完全基因敲除和条件型基因敲除 (又称不完全基因敲除)两种。完全基因敲除是指通过同源重组法完全消除细胞或者动物个体中的靶基因活性,条件型基因敲除是指通过定位重组系统实现特定时间和空间的基因敲除。噬菌体的 Cre/Loxp 系统、Gin/Gix 系统、酵母细胞的 FLP/FRT 系统和 R/RS 系统是现阶段常用的四种定位重组系统, 尤以 Cre/Loxp 系统应用得最

为广泛。

1. 完全基因敲除

实验中一般采用取代型或插入型载体(图 6-9)在 ES 细胞中根据正-负双向选择(positive-negative selection,PNS)原理(图 6-10)进行完全基因敲除实验。正向选择基因 *neo* 通常被插入载体靶 DNA 功能最关键的外显子中，或通过同源重组法置换靶基因的功能区。*Neo* 基因有双重作用,一方面形成靶位点的插入突变,同时可作为正向筛选标记。负向选择基因 *HSV-tk*(human semian virus-thymidine kinase)则被置于目的片段外侧,含有该基因的重组细胞不能在选择培养基上生长。如果细胞中发生了随机重组,负向选择基因就可能被整合到基因组中,导致细胞死亡。由于基因转移的同源重组自然发生率极低，动物的重组概率约为 $10^{-5}$~$10^{-2}$,植物的概率为 $10^{-5}$~$10^{-4}$,即使采用双向选择法也很难保证一次就从众多细胞中筛选出真正发生了同源重组的胚胎干细胞,还得用 PCR 及 Southern 杂交等多种分子筛选技术验证所获得的确实是目的基因被敲除的细胞系。

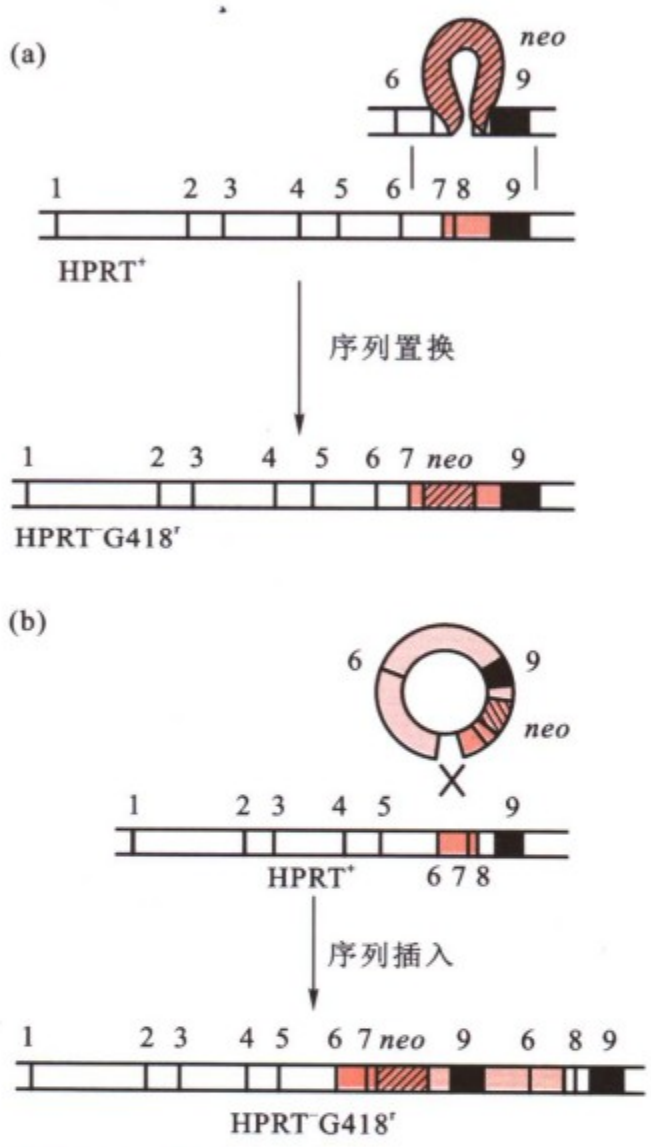



图 6-9 用取代型(a)或插入型(b)载体进行完全基因敲除实验

2. 条件型基因敲除

对于许多有重要生理功能的基因来说，完全基因敲除往往导致胚胎死亡，使有关基因功能的研究无法开展。因此,条件型基因敲除,特别是应用 Cre/LoxP 和 FLP/FRT 系统所开展的组织特异性敲除，由于其具有可调节性而日益受到科学家的重视。构建条件型基因敲除打靶载体时，常将正向选择标记 *neo* 置于靶基因的内含子中，并在靶基因重要功能域两侧内含子中插入方向相同的 LoxP 位点(图 6-11)。当实验中需要消除靶基因活性时，与带有 Cre 重组酶基因的 ES 细胞杂交,Cre 重组酶就能把两个 LoxP 位点中间的 DNA 片段切除,导致靶基因失活。标记基因两侧也常常带有 LoxP 序列,因为许多时候即使标记基因位于内含子中也会阻断靶基因的转录。一旦出现这种情况,可以用 Cre 重组酶表达质粒转染中靶 ES 细胞,通过 LoxP 位点重组将 *neo* 抗性基因删除。

## 6.2.2 高等动物基因敲除技术

真核生物基因敲除的技术路线主要包括构建重组基因载体，用电穿孔、显微注射等方法把重组 DNA 导入胚胎干细胞纯系中,使外源 DNA 与胚胎干细胞基因组中相应部分发生同源重组，将重组载体中的 DNA 序列整合到内源基因组中并得以表达,主要实验流程及筛选步骤如图 6-12 所示。显微注射命中率较高,但技术难度相对大些。电穿孔法命中率比显微注射低,但操作使用方便。胚

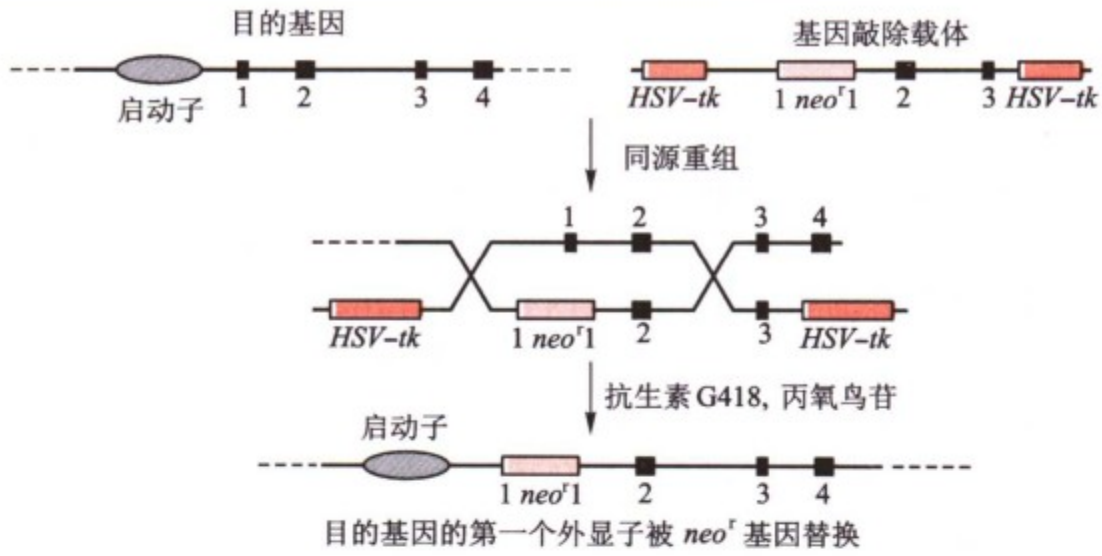



**图 6-10 正负筛选法(PNS 法)筛选已发生同源重组的细胞** 目标序列的第一个外显子中插入 *neo*$^r$ 基因，两侧为单纯疱疹病毒胸腺嘧啶激酶基因，在 ES 细胞中发生同源重组后，*neo*$^r$ 基因插入基因组中，而单纯疱疹病毒胸腺嘧啶激酶基因被剪切掉，发生重组的 ES 细胞由于有 *neo*$^r$ 基因而获得对抗生素 G418 的抗性。因为缺失了单纯疱疹病毒胸腺嘧啶激酶基因，转化细胞同时对嘌呤类似物产生了抗性。

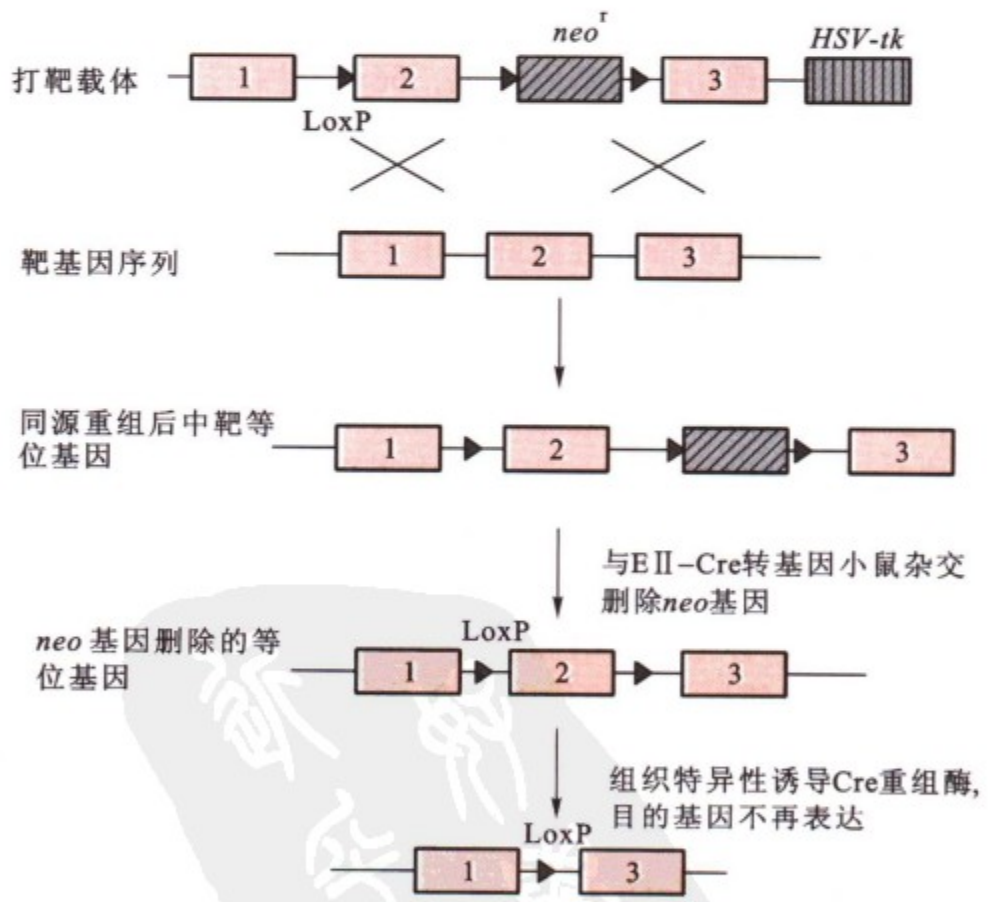



**图 6-11 条件型基因敲除策略** 以 Cre/LoxP 系统为基础，还可以在动物的一定发育阶段和一定组织细胞中实现对特定基因进行遗传修饰的目的。利用控制 Cre 表达的启动子活性或所表达的 Cre 酶活性具有可诱导性的特征，人们常常通过设定诱导时间的方法对动物基因突变的时空特异性进行人为控制，以避免出现死胎或动物出生后不久即死亡的现象。

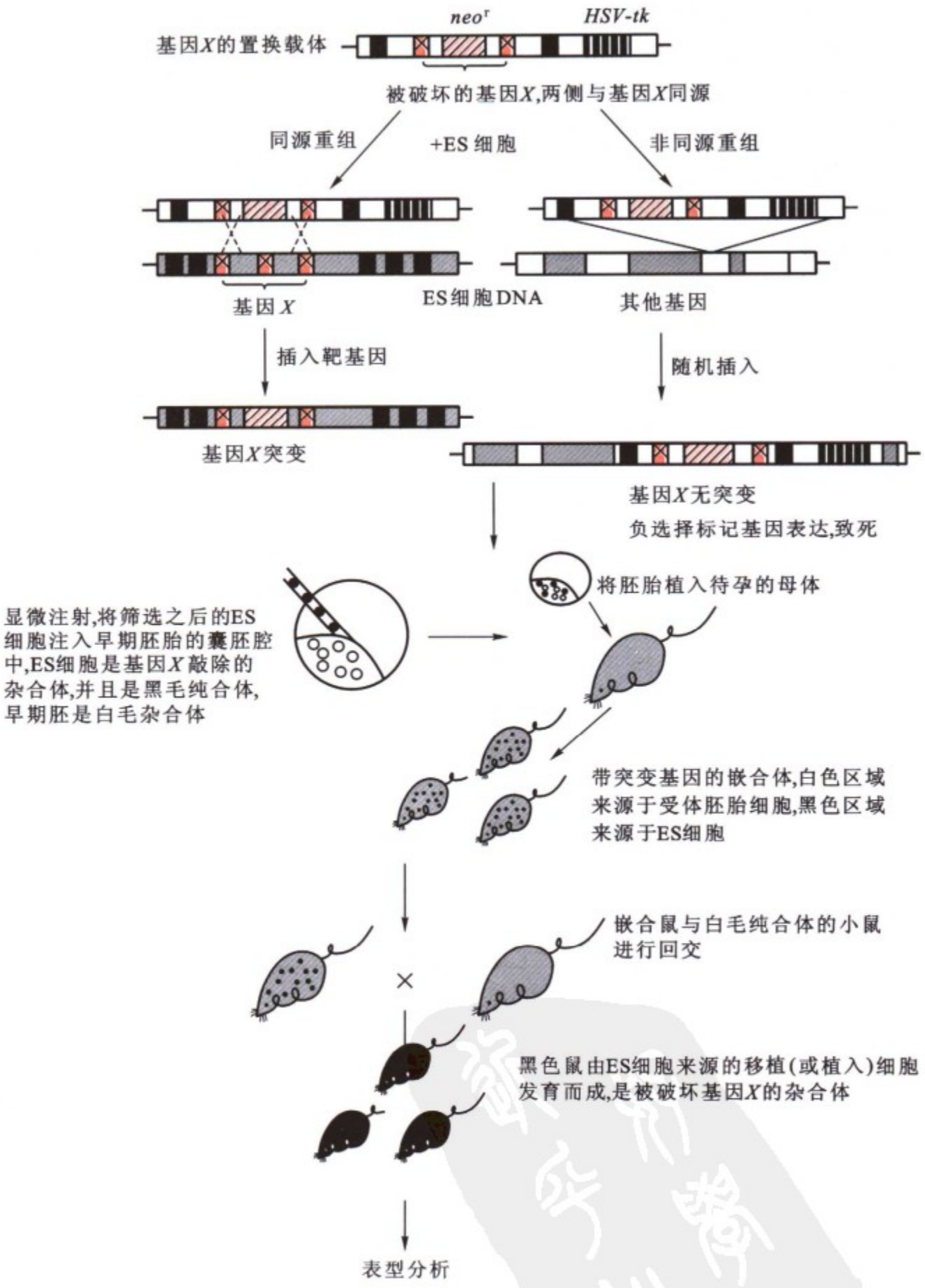



图 6–12 模式动物小鼠中完全基因敲除的主要技术策略与应用举例

胎干细胞(ES 细胞)分离和体外培养的成功奠定了哺乳动物基因敲除的技术基础。目前基因敲除实验中最常用的是小鼠种系 129 及其杂合体,因为这类小鼠具有自发突变成畸胎瘤和畸胎肉瘤的倾向。

因为同源重组常常发生在某一条染色体上,要得到稳定遗传的纯合体基因敲除模型,至少需要两代以上的遗传。除了基因敲除法,还有人用基因捕获的方法通过随机插入突变破坏靶基因表达(图 6-13)。基因捕获载体包括一个无启动子的报道基因(通常为 *neo* 基因),当该基因插入到 ES 细胞染色体某个部位,利用所在位点的转录调控元件得到表达时,该 ES 细胞就获得在含 G418 的选择性培养基上生长的能力。可通过分析标记基因侧翼 cDNA 或染色体 DNA 序列

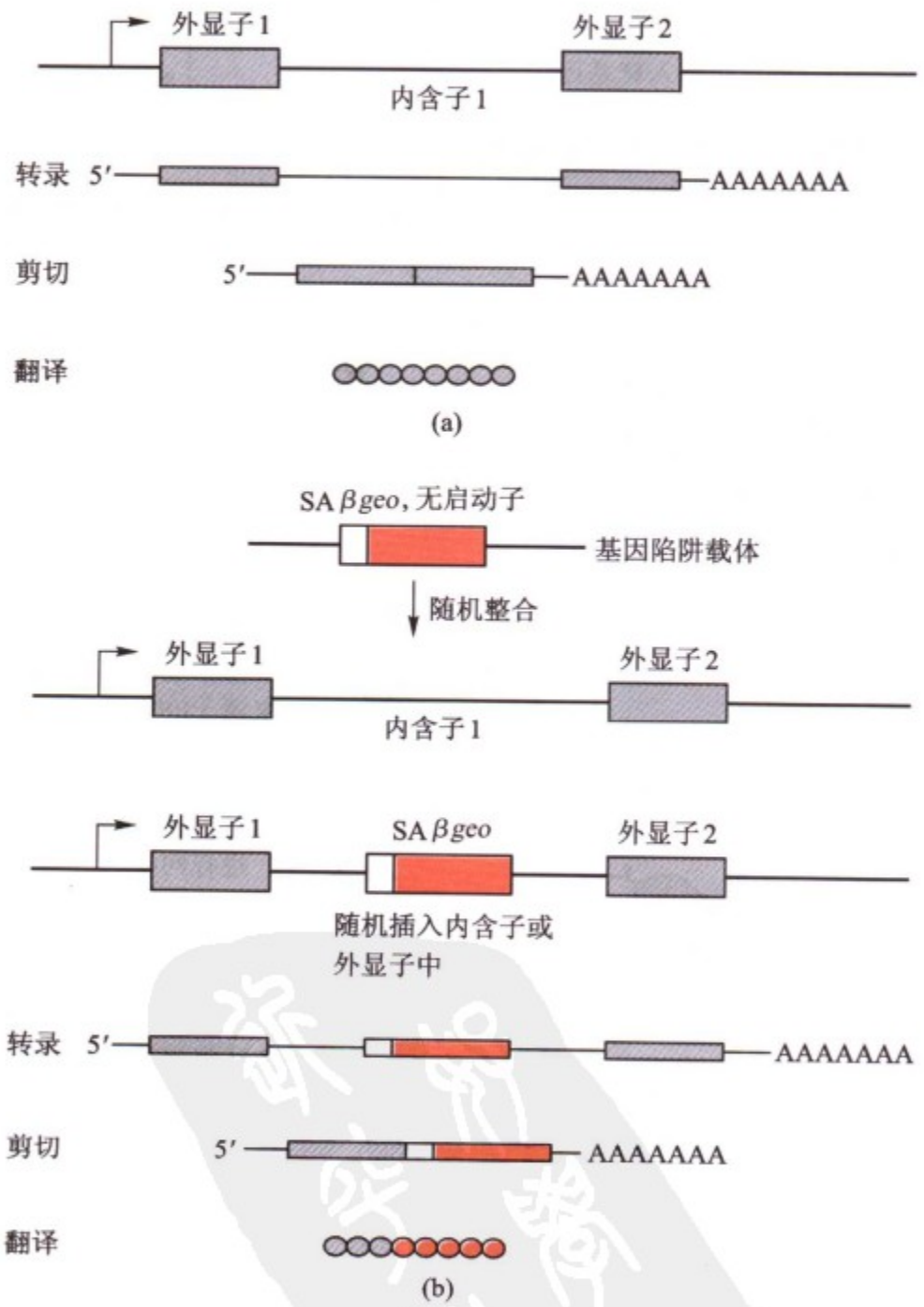



**图 6-13 基因捕获法原理示意图** (a)未插入基因捕获载体前,需要被敲除的靶基因转录翻译出活性蛋白质;(b) 插入基因捕获载体后,靶位点基因转录翻译出带有 *βgeo*(标记基因)蛋白区段的无活性目的蛋白,可用组织化学法检测。

来获得中靶基因的相关信息。

由于受整合位点附近染色质区的影响，转基因整合具有显著的位置效应，基因 5′端启动子和增强子区，3′端终止子区都会对转基因的整合产生影响，这是某些打靶载体选择组织特异性启动子的原因。外源转基因的有效表达有时还取决于其是否有内含子，因为内含子中存在的转录调控元件可影响 mRNA 的剪切以及启动子与内含子间的调节反应。另外，内含子可能含有能开放染色体功能域的序列，还可能通过影响核质成分、位置等来提高转基因表达。除了研究基因功能之外，基因敲除技术还被广泛应用于建立人类疾病的转基因动物模型，为医学研究提供遗传学数据，为遗传病的治疗、生物新品种的培育奠定了良好的基础。

### 6.2.3 植物基因敲除技术

由于植物和动物细胞的差别，基因敲除技术在动物和植物细胞中的操作存在着显著差异。T-DNA 插入失活技术是目前在植物中使用最为广泛的基因敲除手段。T-DNA 插入失活就是利用根癌农杆菌 T-DNA 介导转化，将一段带有报告基因的 DNA 序列标签整合到基因组 DNA 上，如果这段 DNA 插入到目的基因内部或附近，就会影响该基因的表达，从而使该基因“失活”。由于该基因内部或附近插入了一段已知序列的 DNA，可据此设计引物，用 PCR 方法将被破坏的靶基因序列分离出来(图 6-14a)。若将靶基因(假定编码区为 900 bp)两端的引物 LP、RP 及插入载体上的引物 LB 加入同一反应体系中进行 PCR，理论上能得到三种类型的条带(图 6-14b)。野生型植株中，只有 LP 和 RP 引物配对扩增

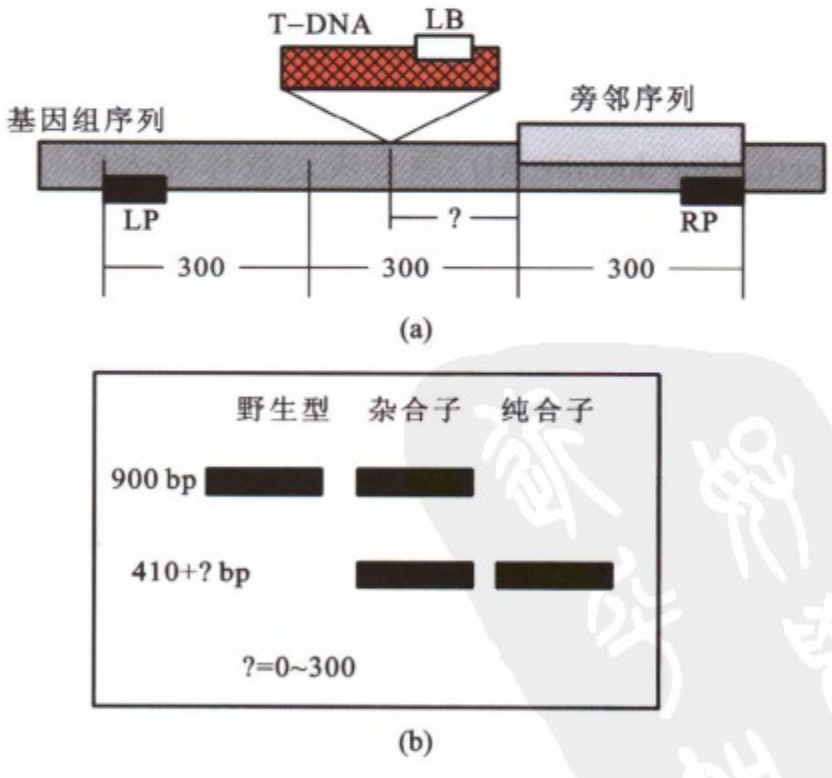



**图 6-14 植物基因敲除突变体植株的筛选** (a)引物设计。LP 和 RP 分别代表插入基因左右两端的引物，LB 是指插入载体上的一段引物，目的基因片段假设为 900 bp，旁邻序列是经测序后获得的 DNA 序列；(b)PCR 产物电泳结果。分别代表野生型、杂合子以及纯合子的PCR 条带。

出来的靶基因条带;如果实验材料来自纯合型基因敲除植株,那么,只有靶基因一端的引物可以与 LB 引物配对完成 PCR 扩增;如果实验材料来自杂合型基因敲除植株,那么,PCR 扩增后会同时出现两种条带。

因为 T-DNA 无专一整合位点,在植物基因组中发生随机整合,所以,只要突变株的数目足够大,从理论上就可能获得任何一个功能基因都发生突变的基因敲除植物库。因为拟南芥基因组冗余序列少,基因密度高,几乎每一个 DNA 插入都会导致某个基因功能的丧失,结合已完成的拟南芥基因组信息,人们很容易筛选到新的功能基因。由于用 T-DNA 插入失活方法建立了拟南芥大规模基因敲除突变体库,研究人员可以方便地在多个数据库中检索到感兴趣基因的突变体,再开展深入的表型分析和功能研究。

## 6.3 蛋白质及 RNA 相互作用技术

### 6.3.1 酵母单杂交系统

酵母单杂交系统(yeast one-hybrid system)是 20 世纪 90 年代中期发展起来的新技术,常用于研究 DNA-蛋白质之间的相互作用。其特点是可识别稳定结合于 DNA 上的蛋白质,可在酵母细胞内研究真核生物中 DNA-蛋白质之间的相互作用,并通过筛选 DNA 文库直接获得靶序列相互作用蛋白的编码基因。此外,该体系也是分析鉴定细胞中转录调控因子与顺式作用元件相互作用的有效方法。

酵母单杂交的基本原理如图 6-15 所示,首先将已知的特定顺式作用元件构建到最基本启动子(minimal promoter,Pmin)的上游,把报告基因连接到 Pmin 下游。然后,将编码待测转录因子 cDNA 与已知酵母转录激活结构域(transcription-activating domain,AD)融合表达载体导入酵母细胞,该基因产物如果能够与顺式作用元件相结合,就能激活 Pmin 启动子,使报告基因得到表达。

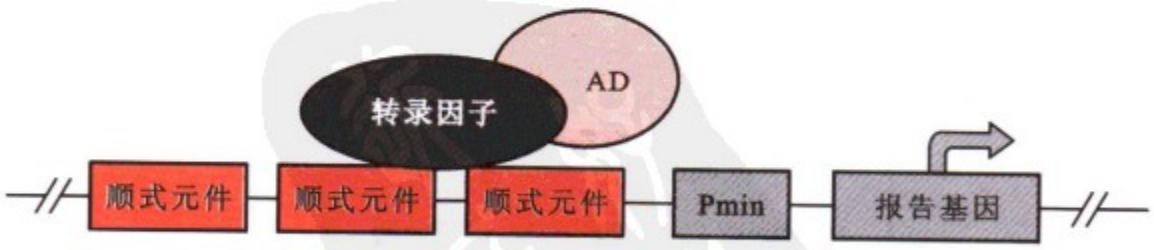



**图 6-15 酵母单杂交的基本原理示意图** 转录因子与顺式元件结合,激活最基本启动子 Pmin,使报告基因表达。若接入 3 个以上顺式作用元件,可增强转录因子的识别和结合效率。

目前,酵母单杂交体系主要被用于确定某个 DNA 分子与某个蛋白质之间是否存在相互作用,被用于分离编码结合于特定顺式调控元件或其他 DNA 位点的功能蛋白编码基因,验证反式转录调控因子的 DNA 结合结构域,准确定位

参与特定蛋白质结合的核苷酸序列。由于该方法的敏感性和可靠性,现已被广泛用于克隆细胞中含量极低且用生物化学手段难以纯化的那部分转录调控因子。常用的酵母单杂交体系基本选用 *HIS*3 或 *LacZ* 作为报告基因,虽然有的体系将带有报告基因的载体直接整合于酵母染色体上,在大部分的实验中报告基因都位于质粒 DNA 上。图 6-16 是利用酵母单杂交体系和已知的顺式作用元件 DRE 从拟南芥 cDNA 文库中筛选转录调控因子的基本流程。实验中如果将不同的未知基因与酵母 GAL4 的 DNA 结合结构域相融合,通过检测位于GAL4 顺式作用元件下游的报告基因的表达状况,还可以鉴定出该转录因子是否具有转录激活功能。

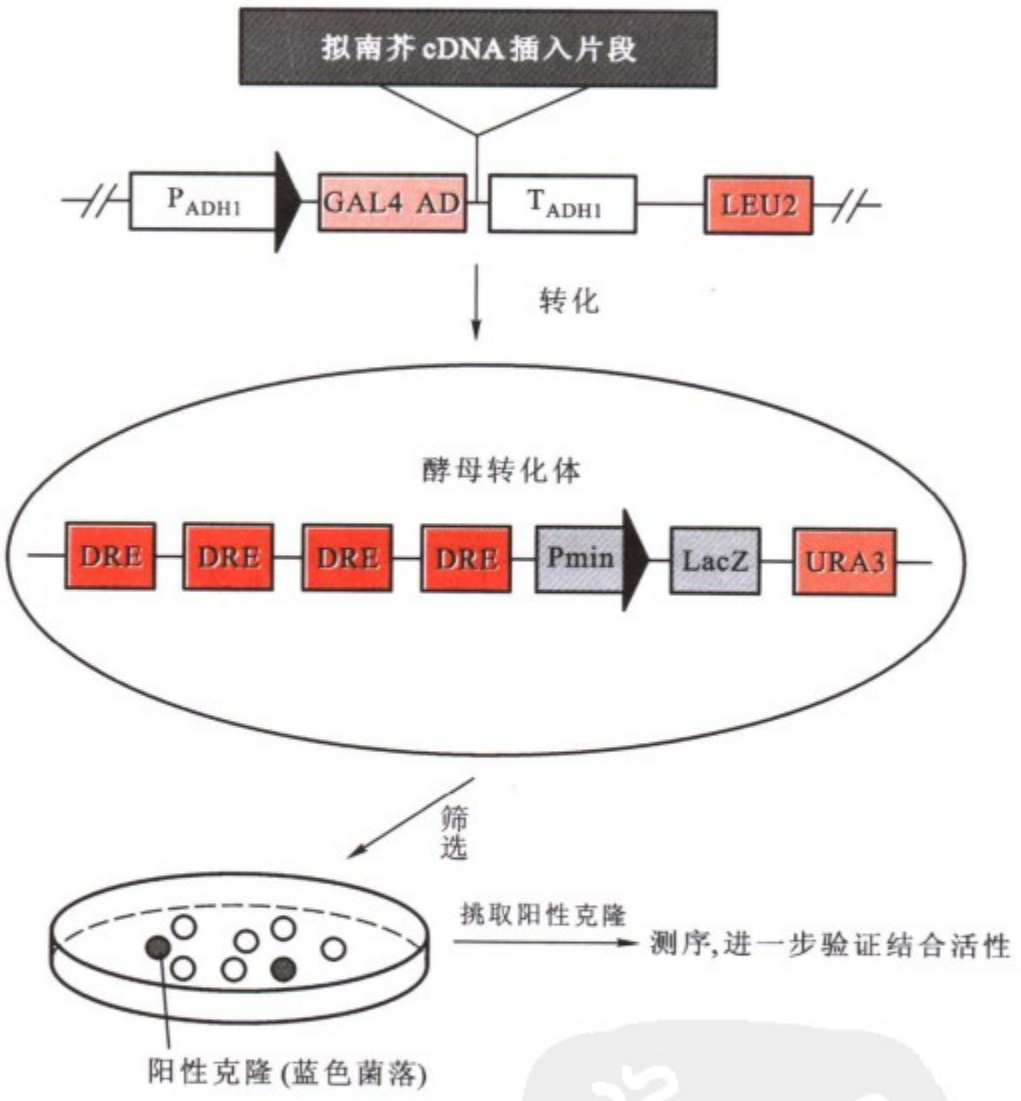



图 6-16 从拟南芥 cDNA 文库中筛选与顺式元件 DRE 结合的转录因子示意图

## 6.3.2 酵母双杂交系统

酵母双杂交体系(yeast two-hybrid system)巧妙地利用真核生物转录调控因子的组件式结构(modular)特征,因为这些蛋白往往由两个或两个以上相互独立的结构域构成,其中 DNA 结合结构域(binding domain,BD)和转录激活结构域(activationg domain,AD)是转录激活因子发挥功能所必需的。单独的 BD 能与特定基因的启动区结合,但不能激活基因的转录,而由不同转录调控因子的 BD 和

AD 所形成的杂合蛋白却能行使激活转录的功能。实验中,首先运用基因重组技术把编码已知蛋白的 DNA 序列连接到带有酵母转录调控因子 (常为 GAL1、GAL4 或 GCN1)DNA 结合结构域编码区(BD)的表达载体上。导入酵母细胞中使之表达带有 DNA 结合结构域的杂合蛋白, 与报告基因上游的启动调控区相结合,准备作为"诱饵"捕获与已知蛋白相互作用的基因产物。此时,若将已知的编码转录激活结构域(AD)的 DNA 分别与待筛选的 cDNA 文库中不同插入片段相连接,获得"猎物"载体,转化含有"诱饵"的酵母细胞。一旦酵母细胞中表达的"诱饵"蛋白与"猎物"载体中表达的某个蛋白质发生相互作用,不同转录调控因子的 AD 和 BD 结构域就会被牵引靠拢,激活报告基因表达。分离有报告基因活性的酵母细胞,得到所需要的"猎物"载体,获得与已知蛋白相互作用的新基因(图 6-17)。

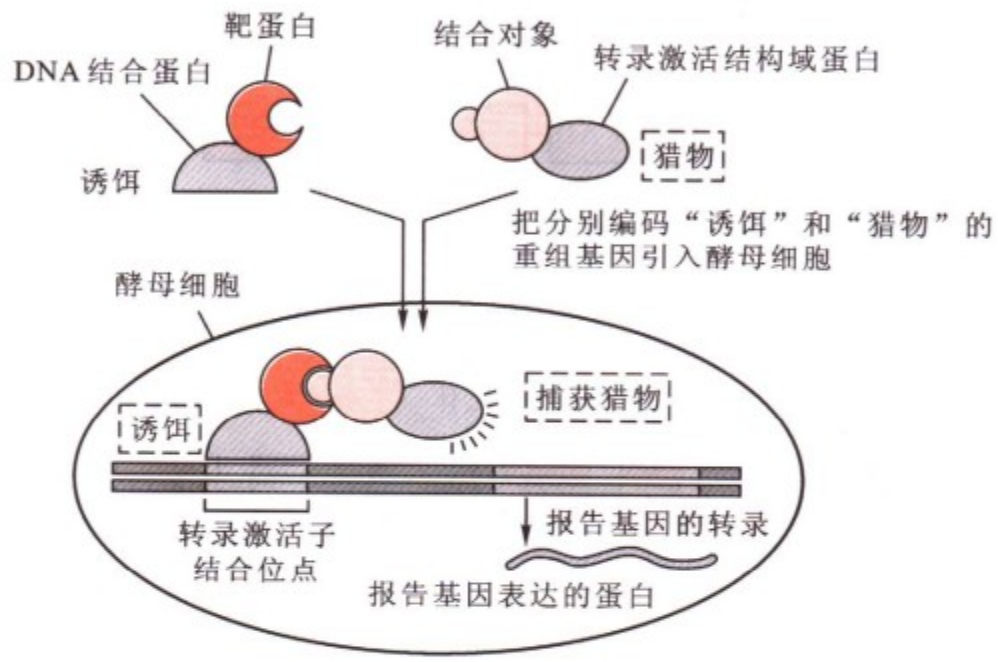



图 6-17 酵母双杂交技术原理示意图

## 6.3.3 体外蛋白质相互作用技术

1. Far Western 印迹技术

典型的 Far Western 印迹试验如图 6-18 所示, 探针蛋白是一种 GST 融合物,包含了蛋白酶切割位点和一种已知蛋白激酶的靶位点。将表达的融合蛋白结合到谷胱甘肽球珠上,用商业性蛋白激酶进行 $^{32}P$ 标记。用蛋白酶切下 GST 寡肽后用谷胱甘肽洗脱被标记的蛋白探针。用 SDS-聚丙烯酰胺凝胶分离待测蛋白提取物并转移到尼龙膜上,也可将表达文库涂板后用吸印法将蛋白质转移到膜上。与探针孵育后,洗膜并进行放射自显影。与标记蛋白发生相互作用的某个或某些条带克隆就是编码与探针蛋白相互作用蛋白的 cDNA。此外,抗 GST 的抗体可以用于定位 Far Western 印迹中非标记融合蛋白, 也可通过组织化学法用与抗生物素抗体相偶联的酶来检测生物素标记的 GST。

融合蛋白的设计依赖于所选择的检测及筛选方式。另外要考虑的是探针蛋白的哪一个区域适合于被嵌入 GST 融合蛋白中。进行文库筛选时,采用尽可能

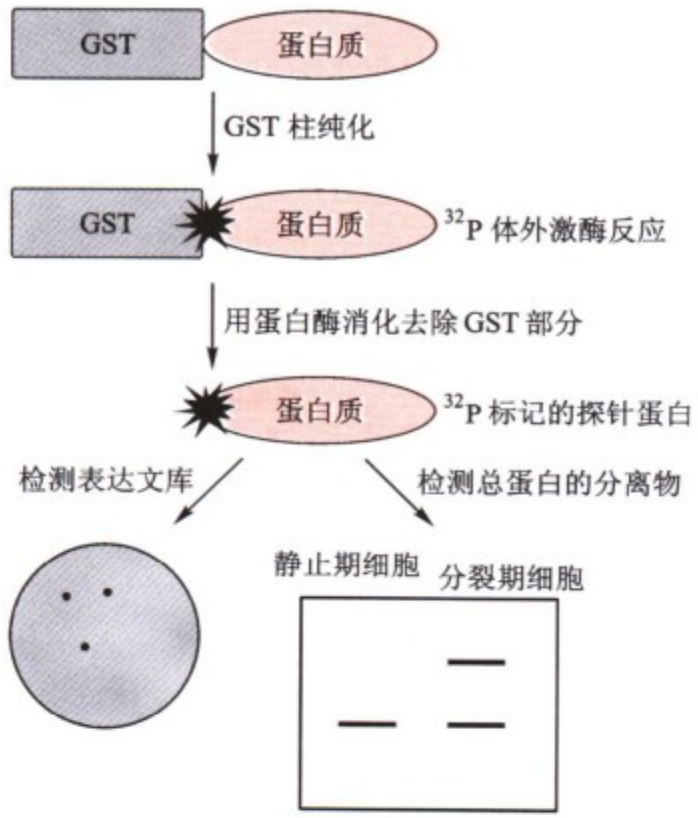



图 6–18 Far Western 印迹试验

大的探针蛋白可以增加可被检测到的相互作用蛋白的数量。如果探针蛋白含有已知的相互作用区并且已被证实具有相互作用能力,那么,采用仅含这些结构区的融合蛋白可能更好。

2. GST 融合蛋白沉降技术

本技术利用 GST 对与谷胱甘肽偶联的琼脂糖球珠的亲和性，从混合蛋白质样品中纯化得到相互作用蛋白(图 6–19)。GST 沉降试验通常有两种应用:确定探针蛋白与未知蛋白间的相互作用,证实探针蛋白与某个已知蛋白之间的相互作用。

3. 蛋白质芯片技术

蛋白质芯片可以用来大规模筛选蛋白之间的相互作用,将一个蛋白芯片与荧光标记的探针蛋白孵育,洗脱非特异性结合的蛋白后,可以通过扫描芯片上的荧光点来检测稳定的相互作用蛋白点(图 6–20)。

4. 等离子表面共振技术

该技术是将诱饵蛋白结合于葡聚糖表面,将葡聚糖层固定于纳米级厚度的金属膜表面。当有蛋白质混合物经过时,如果有蛋白质同“诱饵”蛋白发生相互作用,那么两者的结合将使金属膜表面的折射率上升,从而导致共振角度的改变。而共振角度的改变与该处的蛋白质浓度呈线性关系,由此可以检测蛋白质之间的相互作用(图 6–21)。该技术不需要标志物和染料,安全灵敏快速,还可定量分析。缺点是需要专门的等离子表面共振检测仪器。

5. 免疫共沉淀技术

这一技术的核心是通过抗体来特异性识别候选蛋白。首先,将靶蛋白的抗体通过亲和反应连接到固体基质上，再将可能与靶蛋白发生相互作用的待筛选蛋白加入反应体系中，用低离心力沉淀或微膜过滤法在固体基质和抗

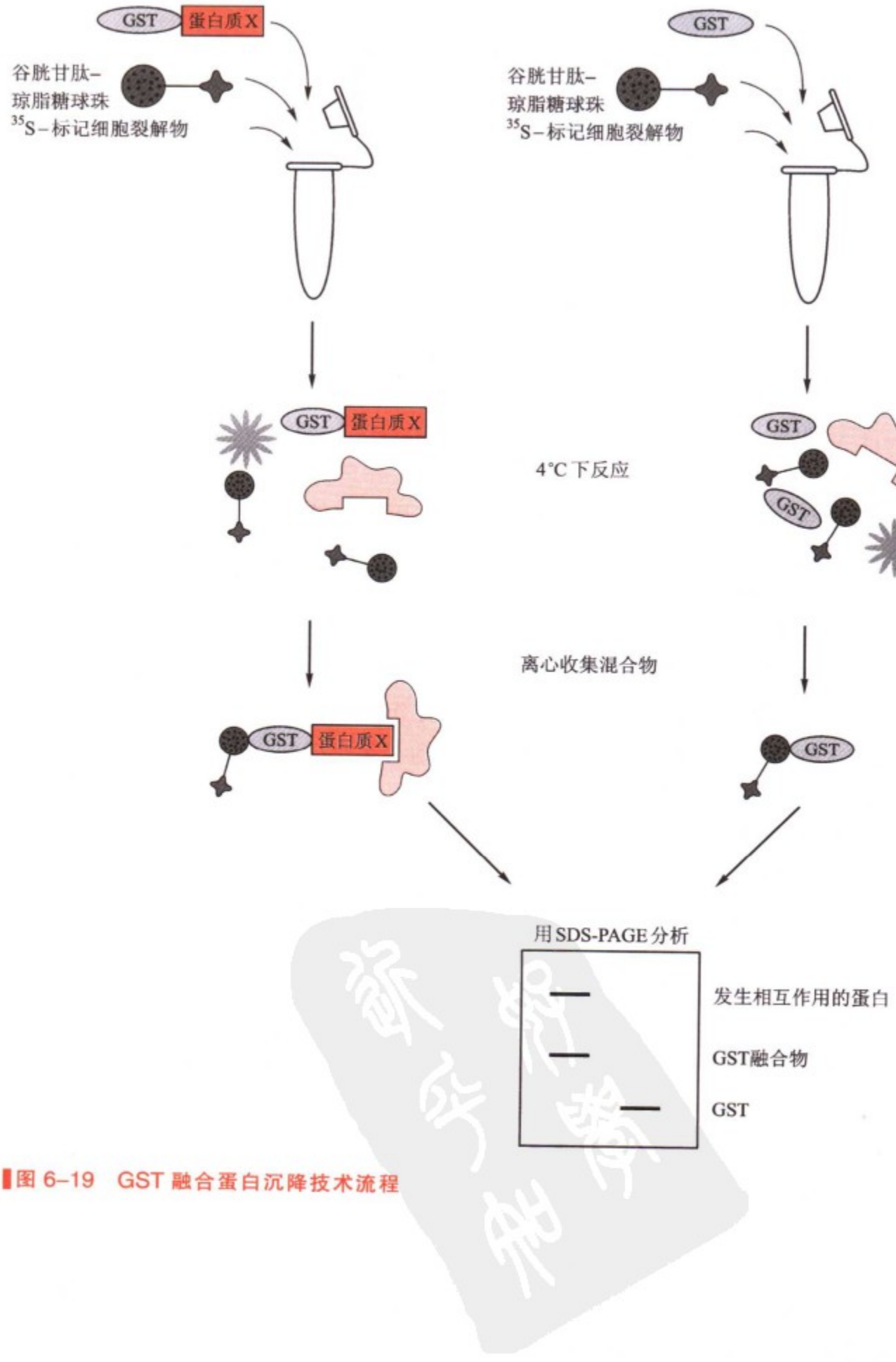



图 6-19 GST 融合蛋白沉降技术流程

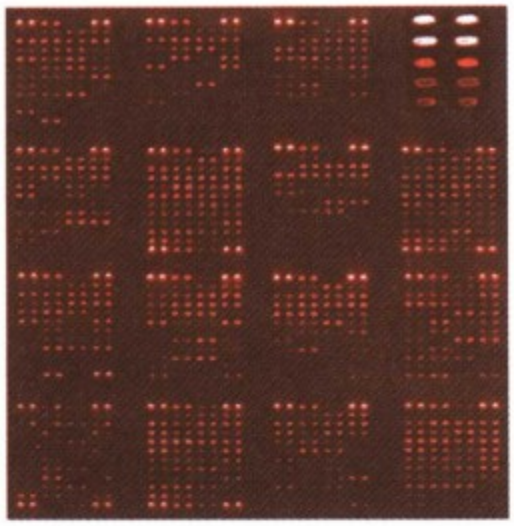


**图 6-20 蛋白质芯片的制备** 将 533 个拟南芥转录调控因子的 HIS 融合蛋白点到玻璃芯片上，以 anti-His 抗体作为一抗检测转录调控因子蛋白质总量。每个蛋白质都重复点样两次。右上角从上到下依次为 1.0、0.1 μg，10、1 和 0.1 ng 的 IgG 标准样品的杂交结果。

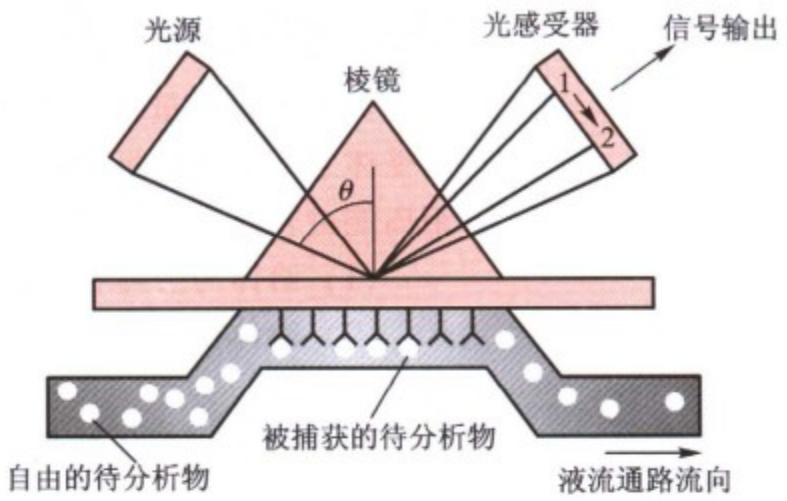



**图 6-21 离子表面共振技术示意图**

体的共同作用下将蛋白复合物沉淀到试管的底部或微膜上。如果靶蛋白与待筛选蛋白发生了相互作用，那么，这个待筛选蛋白质就通过靶蛋白与抗体和固体基质相互作用而被分离出来(图 6-22)。

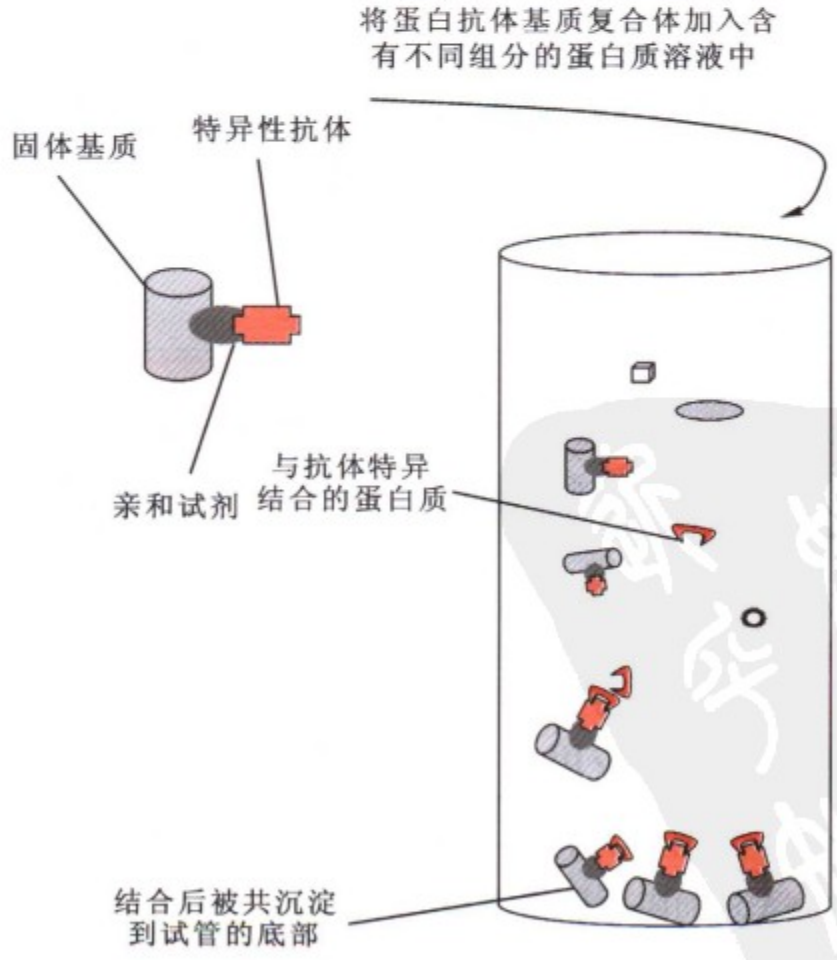



**图 6-22 免疫共沉淀示意图**

### 6.3.4 细胞内蛋白质相互作用研究——荧光共振能量转移法

荧光共振能量转移(FRET)现象是20世纪初发现的。FRET荧光能量给体与受体之间通过偶极-偶极耦合作用以非辐射方式转移能量的过程又被称为长距离能量转移,有三个基本条件:

(1) 给体与受体在合适的距离(1~10 nm)。

(2) 给体的发射光谱与受体的吸收光谱有一定的重叠(这是能量匹配的条件)。

(3) 给体与受体的偶极具有一定的空间取向(这是偶极-偶极耦合作用的条件)。

FRET需要有两个探针,即荧光供体和荧光受体,要求供体的发射光谱与受体的吸收光谱有一定的重叠,而与受体的发射光谱尽量没有重叠。常用的探针有三种:荧光蛋白,传统有机染料和镧系染料。

荧光蛋白是一类能发射荧光的天然蛋白及其突变体,常见的有绿色荧光蛋白(GFP)、蓝色荧光蛋白(BFP)、青色荧光蛋白(CFP)和黄色荧光蛋白(YFP)等。不同蛋白的吸收和发射波长不同,可根据需要组成不同的探针对。传统有机染料是指一些具有特征吸收和发射光谱的有机化合物组成的染料对。常见的有荧光素、罗丹明类化合物和青色染料Cy3、Cy5等,该类染料分子体积较小、种类较多且大部分为商品化的分子探针染料,因此应用较为广泛。镧系染料一般与有机染料联用,分别作为FRET的供体或受体,检测的准确性和信噪比较传统染料有显著提高。

研究蛋白质间相互作用时,FRET一般与荧光成像技术联用,将蛋白标记上荧光探针,当蛋白间不发生作用时,其相对距离较大,无FRET现象,而蛋白发生相互作用时,其相对距离缩小,FRET发生(图6-23),用LSCM或WFM纤维技术,可根据成像照片的色彩变化很直接地反映该过程的发生。

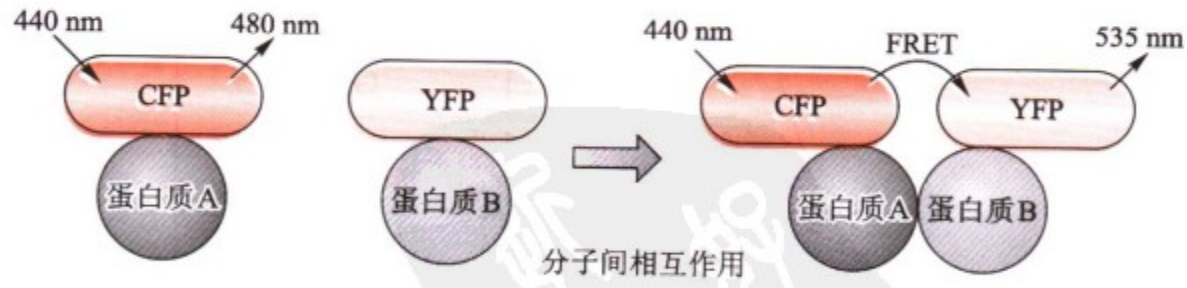



图6-23 荧光共振能量转移法

### 6.3.5 RNAi技术及其应用

RNAi(RNA interference,RNA干涉)技术利用双链小RNA高效、特异性降解细胞内同源mRNA从而阻断靶基因表达,使细胞出现靶基因缺失的表型。RNAi的命名来源于安德鲁·法尔1998年在《自然》杂志上的论文。他们研究发现,对线虫注射外源双链RNA(double-stranded RNA,dsRNA)可诱发与该RNA高度同源

的基因序列的特异性“沉默”。现在已经知道,RNAi 是一个涉及多步通路的过程,包括对触发物的加工、与目标 mRNA 的结合以及目标 mRNA 的降解。至今已在果蝇、锥虫、涡虫、线虫等许多动物及大部分植物中陆续发现了RNAi 效应。

研究发现,双链 RNA 是 RNAi 的触发物,引发与之互补的单链 RNA(single-stranded RNA,ssRNA)的降解。经过 Dicer(一种具有 RNase Ⅲ活性的核酸酶)的加工,细胞中较长的外源双链 RNA(30 个核苷酸以上)首先被降解形成21~25 个核苷酸的小分子干扰核糖核酸(short interfering RNA,siRNA),并有效地定位目标 mRNA。因此,siRNA 是引发转录后基因沉默中序列特异性 RNA 降解的重要中间媒介。较短的双链 RNA 不能被有效地加工为 siRNA,因而不能介导RNAi。siRNA 具有特殊的结构特征,即 5′端磷酸基团和 3′端的羟基,其两条链的 3′端各有两个碱基突出于末端。由 siRNA 中的反义链指导合成一种被称为 RNA 诱导的沉默复合体(RISC)的核糖体,再由 RISC 介导切割目的 mRNA 分子中与 siRNA 反义链互补的区域,从而实现干扰靶基因表达的功能(图 6-24)。同时,

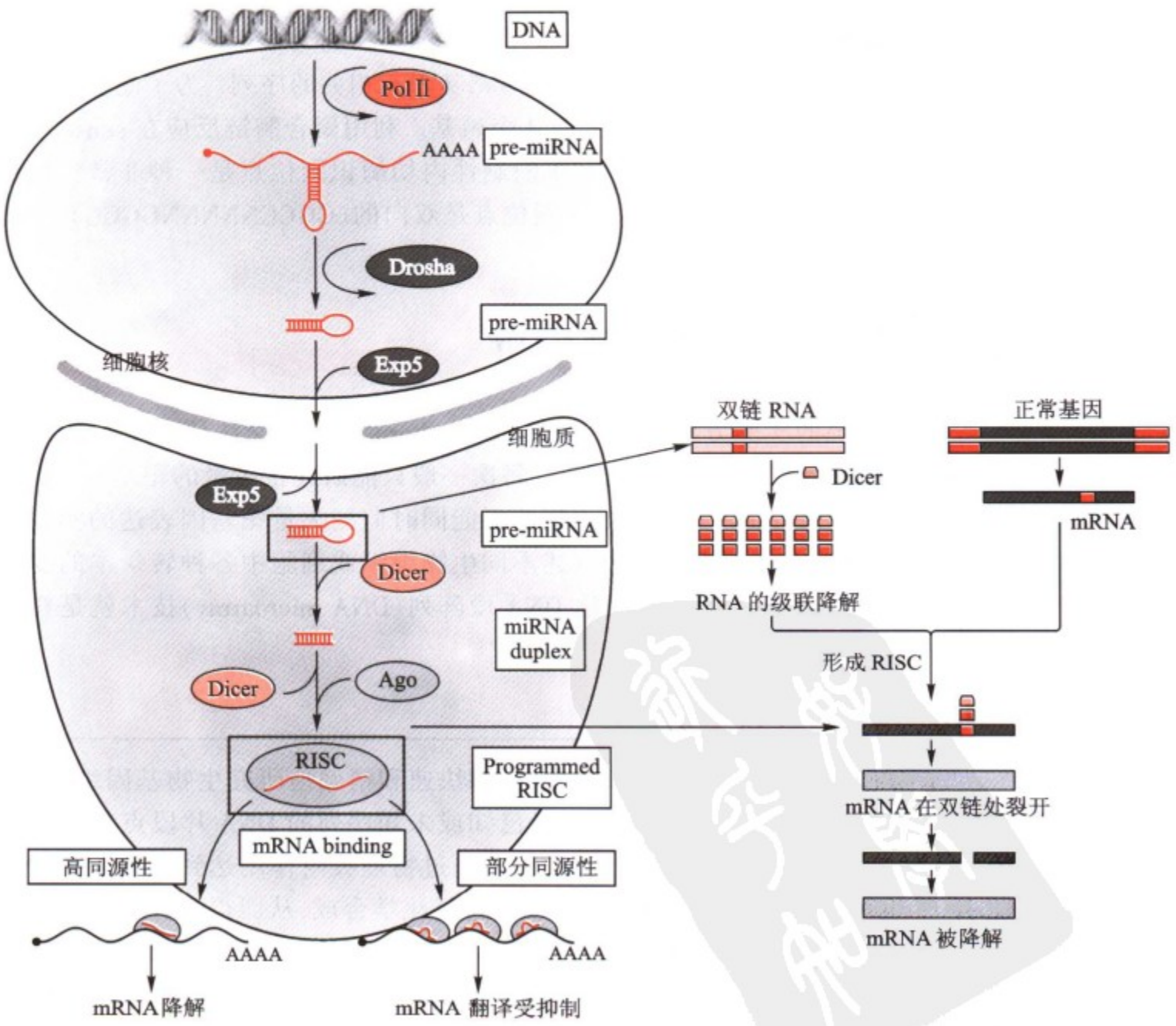



图 6-24 RNAi 作用机制示意图

siRNA 可作为特殊的引物，在依赖于 RNA 的 RNA 聚合酶的作用下，以目的 mRNA 为模板合成 dsRNA,后者又可被降解为新的 siRNA,重新进入上述循环。因此,即使外源 siRNA 的注入量较低,该信号也可能迅速被放大,导致全面的基因沉默。

非哺乳动物细胞可以利用较长的双链 RNA 直接诱导产生 RNAi 而无需合成 siRNA,因此,设计非哺乳动物细胞 RNAi 实验的步骤比较简便。只要通过目的基因体外转录得到所需要的 dsRNA,再通过浸泡、注射或转染靶细胞即可实现 RNAi。在哺乳动物细胞内,较长的 dsRNA 会导致非特异性基因沉默,只有大约 21~25 个核苷酸的 siRNA 才能有效地引发特异性基因沉默。

拟南芥中一般通过以下几种途径产生 dsRNA:将过量表达 sense RNA 的株系与另一个过量表达 antisense RNA 的株系进行杂交,或将带有 sense 和antisense RNA 的 T-DNA 转入同一个植株中,或将具备自主产生双链 RNA(转录产物带有发卡式互补结构)的 T-DNA 转入植株中,均可获得发生特异性基因沉默的后代。

形成双链 RNA 需要有一小段非互补区将互补的两个区域分开,这段非互补区可以是靶基因的一部分也可以是与靶基因不相关的序列。为了能够形成发卡(hairpin)结构,非互补区至少要有 4 个碱基。利用聚合酶链反应在 sense 区的 3′端和 antisense 区的 5′端引入 *Sfi* Ⅰ限制性内切酶识别位点是一种非常方便的形成发夹结构的方法,因为 *Sfi* Ⅰ识别位点是双向的(GGCCNNNNNGGCC),中心区域不需要回文序列。

## 6.4 基因芯片及数据分析

若用传统的实验方法如 RT-PCR 或 Northern 印迹杂交法研究基因的表达调控规律,将受电泳泳道数量的限制,每次一般只能研究很少量的靶基因。随着功能基因组学研究的不断深入,迫切需要能同时监测大量靶基因表达的实验手段,迅速准确地在基因组水平上阐述不同生物组织或细胞中各种转录本的变化规律。基因芯片(gene chip),又称 DNA 微阵列(DNA microarray)技术就是在这种情况下应运而生的。

### 6.4.1 基因芯片技术原理

基因芯片是一种小型分析装置，能够快速和精确地研究生物基因组信息。制作基因芯片时，可用机械臂把大量已知或未知序列的 DNA 片段点在玻璃片(通常为 2×2 $cm^2$)、金属片或尼龙膜上,再经过物理吸附作用达到固定化(cDNA 芯片)。也可以直接在玻璃板或金属表面进行化学合成,从而得到寡聚核苷酸芯片。将芯片与待研究的 cDNA 或其他样品杂交,经过计算机扫描和数据处理,便可以观察到成千上万个基因在不同组织或同一组织不同发育时期或不同生理条件下的表达调控情况 (图 6-25)。荧光标记的 cDNA 与芯片上相匹配的DNA 序列发生杂交反应,使得芯片上的点呈现出荧光信号,荧光信号的强度和基因

表达的多度呈正相关。基因芯片这种微型化装置具有巨大的容量,使科学家在单次实验中就可以分析整个基因组的变化。基因芯片分析包括五个基本步骤:生物学问题、样品制备、生物化学反应、检测、数据模型分析。

## 6.4.2 基因芯片的点制过程

1. 简易基因芯片

制作简易基因芯片时,使用机械臂把大量已知或未知序列的 DNA 片段点在一张很小的玻璃片(通常为 2×2 $cm^2$)或尼龙膜上,再经过物理吸附作用达到固定化(cDNA 芯片)。也可以直接在玻璃板表面进行化学合成,从而得到寡聚核苷酸芯片。将芯片与待研究的 cDNA 或其他样品杂交,经过计算机扫描和数据处理,便可以观察到大规模的基因群在不同组织、同一组织不同发育时期或不同生理条件下的表达调控情况(图 6–26)。简易芯片一般的基因数量少(一般不超过 2 000),主要用于研究小部分特定基因的表达状态。简易芯片一般可以用手工制作或者机械手点样。

2. 大规模基因芯片

大规模基因芯片通常涉及基因组规模与数量(一般大于 10 000 个基因),分别采用接触式点样、非接触式点样或者半导体技术制备样品。接触式点样方法包括点样元件或点样元件中的液体和芯片基片间有直接接触的各种方法。非接触式芯片制备方法包括各种芯片制备中不需点样元件和芯片基片间直接接触的各种技术,其样头的核心是以可控的方式从喷嘴中喷出小体积的液滴。半导体制备技术指借鉴计算机芯片制造业中的微型化加工方法的基因芯片制备技术。基因芯片的工作流程如下:

(1) 经大规模 PCR 扩增获得独立 cDNA 插入片段。
(2) 用机械手将 PCR 产物点在玻璃板上,变性、固定。
(3) 分别从不同器官或组织中分离 mRNA,反转录生成 cDNA。
(4) 用不同的荧光染料标记不同器官或组织的 cDNA。
(5) 将标记后的 cDNA 与点好的芯片进行杂交。
(6) 激光扫描芯片杂交结果,计算机处理。
(7) 分析杂交数据。

3. 微珠芯片技术

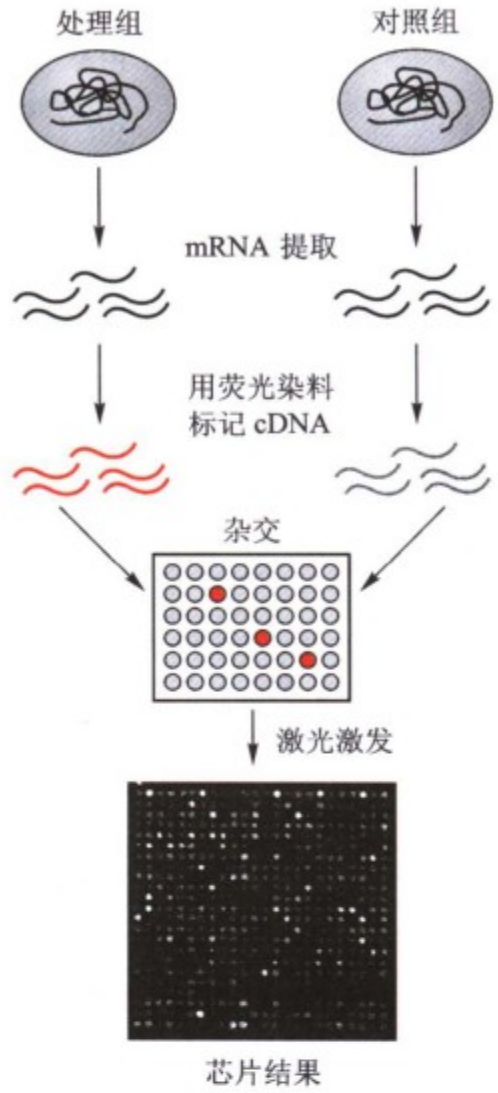



**图 6–25 基因芯片技术流程图** 在基因芯片分析中,首先将从细胞中抽提得到 mRNA 进行荧光标记,与预先制备的含有一定数量和质量 DNA 的尼龙膜、玻璃或金属芯片进行杂交。芯片上某些点能与杂交液中荧光标记的特异性 cDNA 发生杂交结合,所发出的荧光信号与该基因表达的多度呈正相关。通过比较芯片上的每个位置的荧光信号强度来确定基因多度或表达强度。

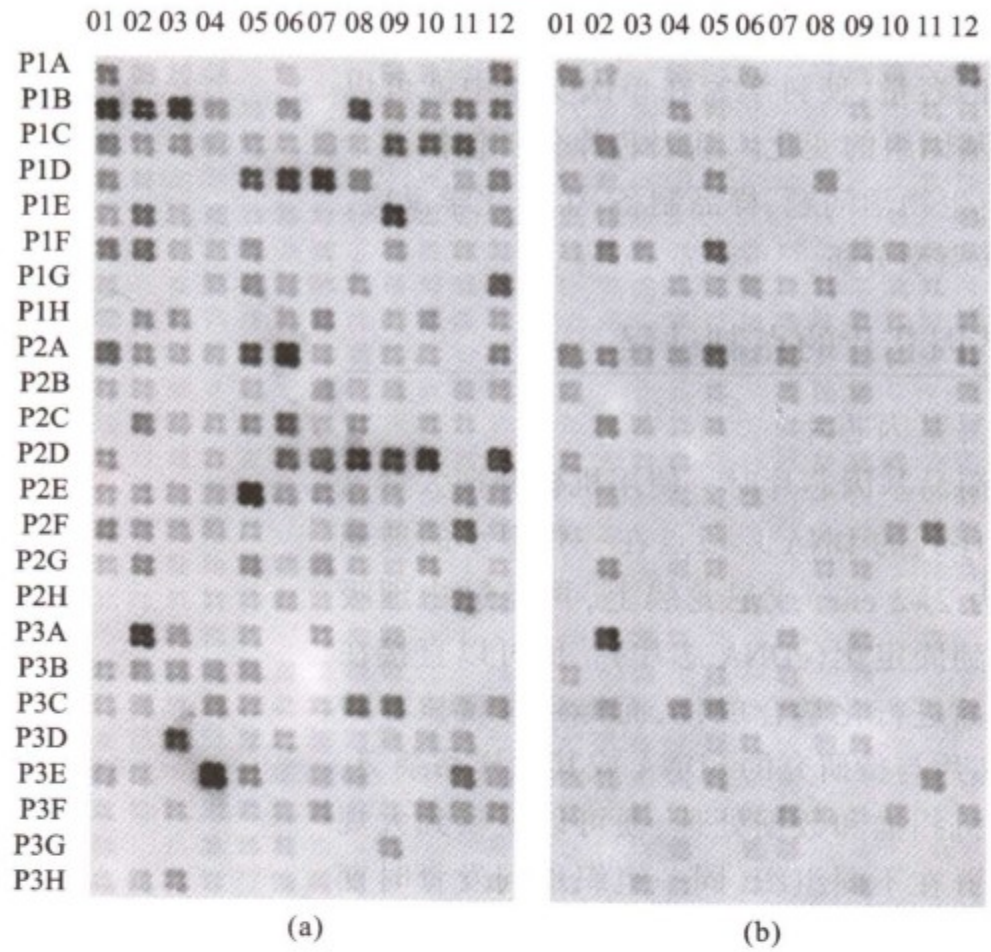



**图 6-26 包含 270 个棉花 cDNA 的简易芯片的杂交图谱** (a)为纤维细胞组织 RNA 杂交结果;(b)为胚珠细胞组织 RNA 杂交结果。

微珠芯片(beadarray)技术是一种新的基因芯片技术,有可能成为今后基因芯片技术发展的方向。目前,微珠芯片技术在生命科学领域的用途主要集中在 SNP 及基因型分析、基因表达谱分析和蛋白组学研究三大领域。微珠芯片技术是在光导纤维技术基础上发展起来的技术。在直径为 3.5 mm 的光纤束中,包含约 50 000 根光纤。在每根光纤的顶部蚀刻出一个小洞,可镶嵌直径为 3 μm 的微磁珠。每个微磁珠上可以根据不同的要求连接不同的配体,如寡核苷酸、蛋白质、抗体等。制作芯片时,将带有所需配体的不同微珠混合后与光纤束进行自组装,通过微珠上寡核苷酸标签序列(address)确认每根光纤上所吸附的不同的微珠(图 6-27)。现有微珠芯片上每束光纤可镶嵌 1 536 种微珠,因此,每种微珠在一束光纤中有 30 次重复。结合在微珠上的寡核苷酸或者蛋白可与荧光标记的配对核酸或配体发生相互作用,发生的信号被激发后通过光纤传递到检测器。

与其他芯片制作技术相比较,微珠芯片具有如下优势:

(1) 密度高 微珠芯片的点样密度是原位光合成法的 16 倍,喷点法的 100 倍,接触式点样的 400 倍。

(2) 测试重复性好 鉴于超高密度的特点,微珠芯片中每个样品都能保证约 30 个重复,从而保证测试的高重复性、高成功率以及高可靠性。

(3) 定制方便 若需要在已完成的芯片中增加测试点或新基因,只要合成相应的微珠加入到微珠混合池中即可。

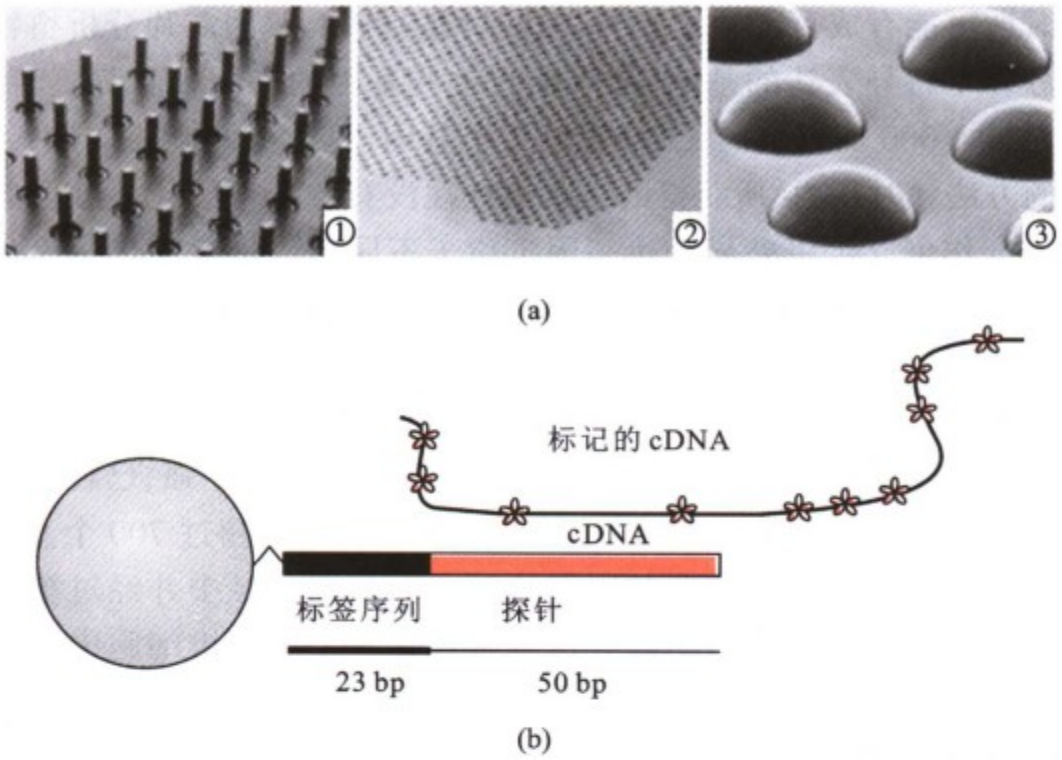



**图 6-27 微珠芯片及原理** (a)微珠芯片制作示意图,①成品芯片;②单束光纤显微截面;③放大后的微磁珠。(b)微珠芯片工作原理。

## 6.4.3 基因芯片数据分析

基因芯片数据分析包括两部分,数据可靠性分析和生物学意义分析。

1. 数据可靠性分析

因为用基因芯片进行杂交分析,每次实验结果都会有些变化,因为包括靶 DNA 浓度、探针浓度、杂交双方的序列组成、盐浓度以及温度等许多因素影响了杂合双链的形成和稳定性。因此,需要进行重复试验以确保数据的准确。主要的方法是通过两次平行杂交反应,将每个基因在两个反应中的表达水平做成对应散点图,只要绝大多数基因的杂交结果都在 45°斜线附近,就说明实验结果是可靠的(图 6-28)。

为了保证基因芯片数据的可靠性,芯片中还需要加上多个持家基因 (housekeeping genes)作为内对照。持家基因,如呼吸作用的酶类和细胞骨架蛋白基因等,是维持细胞正常生长发育的必需基因,其表达水平在细胞内基本不变。处理数据时,认为持家基因的反应信号在两种组织中不变,依此确定其他基因表达信号的相对值。

2. 生物学意义分析

主要指通过分析芯片杂交数据,研究差异表达基因的生物学意义。通常,在芯片中某一器官

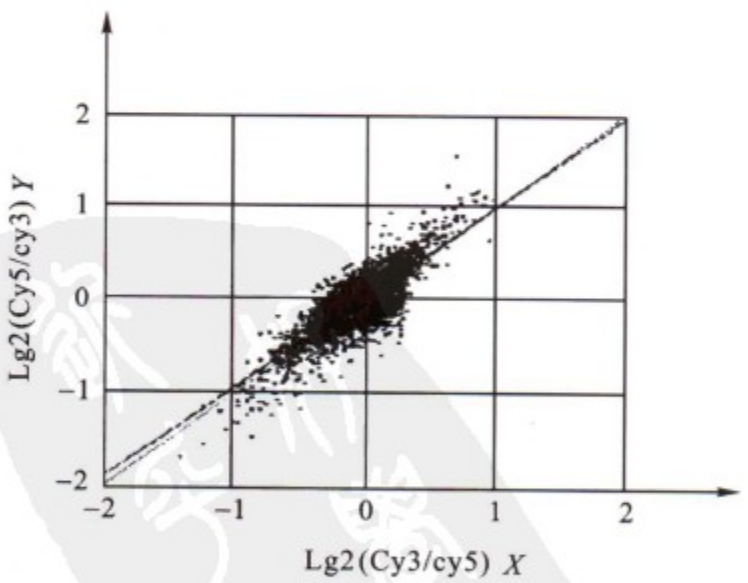



**图 6-28 基因芯片数据可靠性分析** 横坐标 $X$ 表示每个杂交点 Cy3/Cy5 荧光信号的比值,纵坐标 $Y$ 上的点是交换荧光标记后的信号强度比值。若芯片中绝大多数点的信号值在斜线 $Y=X$ 附近,表明该杂交结果可信度较高;反之,则不可信。

或发育时期可能有成百上千个差异表达基因。例如,基因芯片分析植物根部可能有 400 多个特异表达的基因,如果要将这些基因的来龙去脉都搞清楚,可能要追溯超过 4 000 篇以上的文献(假设 1 个基因需要查阅 10 篇文献)。这种不捡重点的方法耗时耗力,所获得的结果也往往没有意义。因此,生物学家和生物信息学家合作创立了将差异表达基因定位于不同的生物化学代谢途径的分析方法(pathway identification)。统计每条生物化学代谢途径中固有的基因数量和被基因芯片定位的差异表达的基因数量,可以计算出特定器官或特定发育阶段表达有显著差异的代谢途径。由于代谢途径通常负责执行生物体的重要功能,并且相对于原始数据,代谢途径的数量相对较少,可操作性强,已经成为目前学术界分析基因芯片数据的重要手段。作者的研究发现有超过 700 个基因在棉花纤维细胞中特异表达,将这些基因定位到不同代谢途径中并经过统计分析发现,乙烯合成、细胞骨架和脂肪酸合成及延伸是三个在纤维细胞中特异性表达的代谢途径(表 6-1)。在后续生物学研究中,以这些最具代表性的代谢途径为突破口,有可能阐明这些途径在纤维细胞发育中的重要性。

表 6-1 用途径分析法研究基因芯片数据中可能具有生物学意义的代谢途径

| 生物体代谢途径 | 芯片中总基因数 | 差异表达基因数 | *P* 值 | *Q* 值 |
|---|---|---|---|---|
| 乙烯合成 | 5 | 3 | 0.0016 | 0.0295 |
| 细胞骨架 | 75 | 10 | 0.0077 | 0.0376 |
| 脂肪酸合成及延伸 | 64 | 9 | 0.0080 | 0.0376 |
| 糖胺降解 | 16 | 4 | 0.0099 | 0.0376 |
| 维生素 C 代谢 | 63 | 8 | 0.0217 | 0.0522 |
| 油菜素内脂合成 | 6 | 2 | 0.0397 | 0.0629 |

## 6.5 利用酵母鉴定靶基因功能

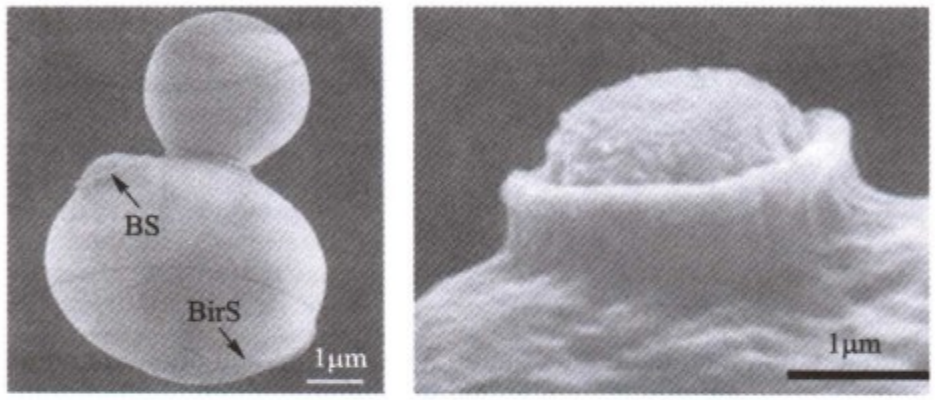
酿酒酵母作为单细胞真核生物,具有和动植物细胞相似的结构特征,包括细胞核、内质网、高尔基体、线粒体、过氧化物酶体、细胞骨架等,而且其细胞生长发育所发生的细胞过程和动植物细胞也有很高的相似性,很多基因在酵母和动植物细胞中是高度保守的。而且,酵母细胞很容易进行生物化学和分子生物学操作,所以,酵母是基因功能研究中常用的模式生物。

### 6.5.1 酵母的遗传学和分子生物学简介

酿酒酵母(*Saccharomyces cerevisiae*)有稳定的单倍体和二倍体细胞,是最早完成基因组 DNA 全序列测定的真核生物。由于其具有快速生长能力、无性繁殖能力以及简便的平板影印能力,已经成为生命科学领域里广泛应用的模式生物。理论上,与任何一种遗传学特征相对应的不同物种的结构基因都可以通过

质粒文库的互补作用而在酵母缺失突变体中得到鉴定。导入酵母细胞中的DNA既能以自主复制的分子形式存在,也可能以被整合到基因组的方式存在。酵母中转化 DNA 的整合重组主要通过同源重组方式进行,整合效率较高。

酵母细胞的直径大约有 5 μm，在光学显微镜下可以观察它的许多重要特征。酿酒酵母细胞是以出芽方式生长的。出芽的细胞被称为母细胞,产生的芽就成为子细胞(图 6-29)。新生的芽从酵母细胞分裂周期一开始就出现,直到细胞周期结束才从母细胞上分离下来。由于酵母细胞的全部生长都集中在芽上,而且这种生长贯穿整个细胞周期,所以通过观察芽的大小就可以粗略估计出一个给定细胞处于哪个生长发育阶段。当培养基中营养元素耗尽时,酵母细胞会停止生长而滞留在某个生长发育阶段,因此,可以通过观察显微镜下酵母细胞的出芽频率确定酵母的营养状态。




(a) (b)

**图 6-29 扫描电镜下的酵母细胞** (a) 新芽常在已脱落的芽斑附近长出来;BS(bud scar),芽痕;BirS(birth scar),蒂痕。(b) 子细胞脱落后的芽斑放大图。

酵母单倍体和二倍体细胞在形态上有相似性,但也存在重大差异。首先,二倍体细胞的直径大约是单倍体的 1.3 倍。其次,二倍体细胞呈长形或卵圆形而单倍体细胞通常接近圆形。第三,二倍体细胞和单倍体细胞的出芽方式不同。每个酵母细胞衰老前一般出芽 20 次,在母细胞表面以一定的方式连续出芽。单倍体细胞按照沿轴线的方向出芽,新芽与前一个芽紧密相连。二倍体细胞按极性方向出芽,新芽可以出现在长形细胞的任何一端。

酵母细胞有三种交配型 MATa,MATα 以及由这两种细胞相互交配形成的第三种交配型 MATa/α。MATa/α 不能与 MATa 和 MATα 中任何一种细胞交配。一般而言,MATa 和 MATα 为单倍体,MATa/α 为二倍体,但也有例外。带有不同突变的单倍体细胞之间发生交配可以将不同的突变基因组合在同一个细胞中,为研究基因之间的相互关系提供了良好的实验材料。

野生型酵母属于原养型,只要培养基中有足够可利用的碳源和氮源,酵母细胞就能合成自身需要的各种营养物质。酵母首选的碳源是葡萄糖,但是也能利用其他很多种糖类。如果碳源适宜,酵母细胞首选通过乙醇发酵产生能量。酵母细胞是条件性厌氧菌,在有氧和无氧状态下都能生存。

二倍体酵母细胞在低氮、无发酵性碳源的条件下(营养缺乏,“饥饿”条件)能通过减数分裂形成单倍体的孢子。此时,每个单倍体细胞在遗传背景上完全相同,从单个孢子来源的所有细胞都带有一套相同的染色体 DNA。这些单倍体细胞中的某些重要基因发生突变后,因为不存在等位基因,很容易导致酵母细胞生长发育异常而发生表型变化。图 6-30 以酵母亮氨酸和色氨酸合成基因为例介绍了二倍体细胞通过减数分裂形成单倍体细胞和发生表型变化的过程。

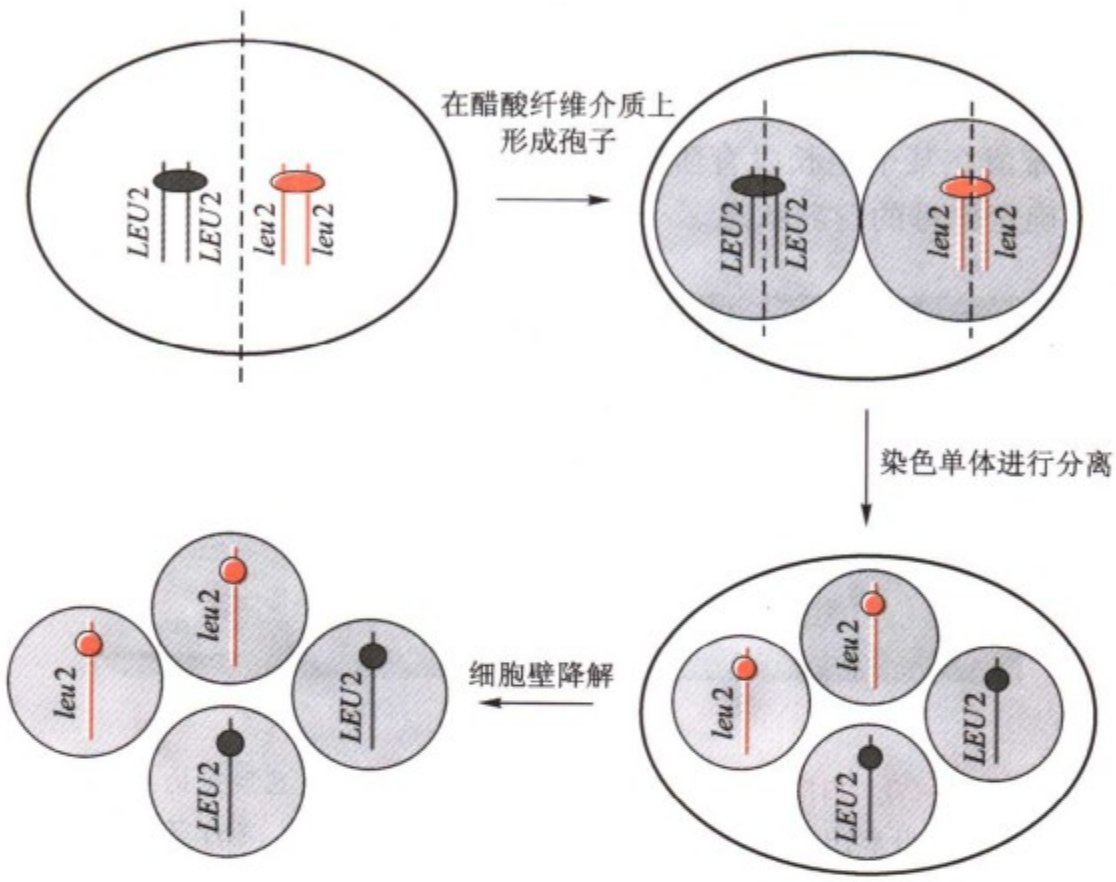



**图 6-30 以亮氨酸合成酶基因为例介绍二倍体细胞通过减数分裂形成单倍体细胞和发生表型变化的过程** 野生型和突变体亮氨酸合成基因(*LEU2*/*leu2*)在酵母细胞形成四个孢子过程中的分配情况。最后形成的四个孢子中有两个孢子带有野生型基因,两个孢子带有突变体基因。

## 6.5.2 酵母基因转化与性状互补

利用 Yip(整合型质粒)可以对酵母基因组中的任意基因进行精确的敲除(置换),并通过孢子繁殖中的四分体分析技术进行观测和研究。带有致死突变位点和可能的外源功能互补基因的酵母二倍体细胞在饥饿条件下形成四分体孢子,将这四个孢子用 SINGER 的酵母四分体分离系统分开,如果所引入的外源基因具有互补突变位点的功能,含有突变基因的单倍体可以存活,否则就不能存活。图 6-31 中将多个棉花 *KCS* 基因(编码超长链脂肪酸合成酶)克隆到含有 URA3 筛选标记的表达质粒上,用带有 URA3 筛选标记的表达质粒替换带有 TRP1 筛选标记的质粒,进一步验证。

基因置换一般在二倍体中进行,因为单倍体基因缺失突变体可能根本无法生存。如果某个基因是生存必需的,四分体孢子的分离比为 2:2;如果这个基因

不是生长所必需的,所有的四分体都能存活,其中有两个孢子属于缺失突变,另两个孢子没有发生变化。酵母基因置换有一步置换法和两步置换法。一步法基因置换是利用包含有突变等位基因以及与靶位点基因末端同源序列来实现的,常用含 Geneticin(G418)的选择性培养基筛选转化子,再用 PCR 方法验证基因是否整合到正确的位置上。用这种方法只需要 2~3 周就可以得到某个特定基因的突变体菌株(图 6-32)。

两步置换法首先要整合一个含有突变等位基因的质粒,目标位点发生整合后,产生了这一基因区段的重复,其中一个为野生型,一个为突变型,两者之间还带有一段质粒 DNA 序列(图 6-33)。两个基本重复拷贝之间随机发生的同源双交换将导致质粒 DNA 的切除以及重复区段中一个拷贝的丢失(留下突变型或野生型拷贝的概率分别是 50%),可通过表型分析、Southern 杂交或 PCR 反应加以区分。质粒上的选择性标记

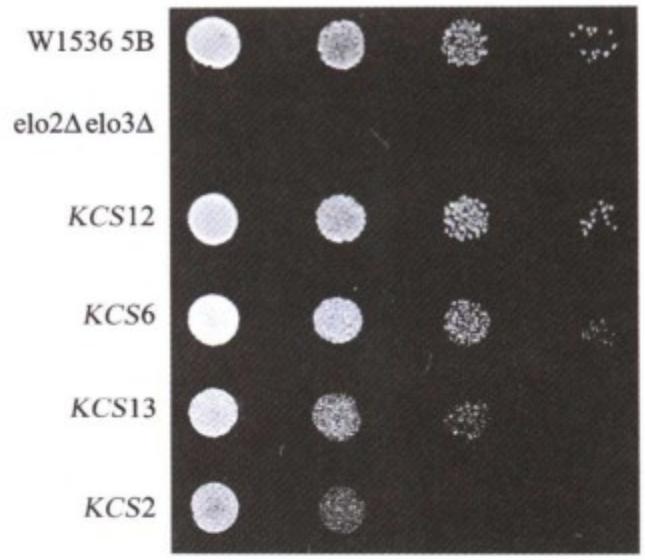



**图 6-31 棉花 *KCS* 基因酵母细胞互补情况** W1536 5B 为野生型单倍体细胞;*elo2Δelo3Δ* 为酵母自身的超长链脂肪酸合成酶双突变的单倍体,为致死型;*KCS12*,*KCS6*,*KCS2*,*KCS13* 为不同的棉花*KCS* 基因。

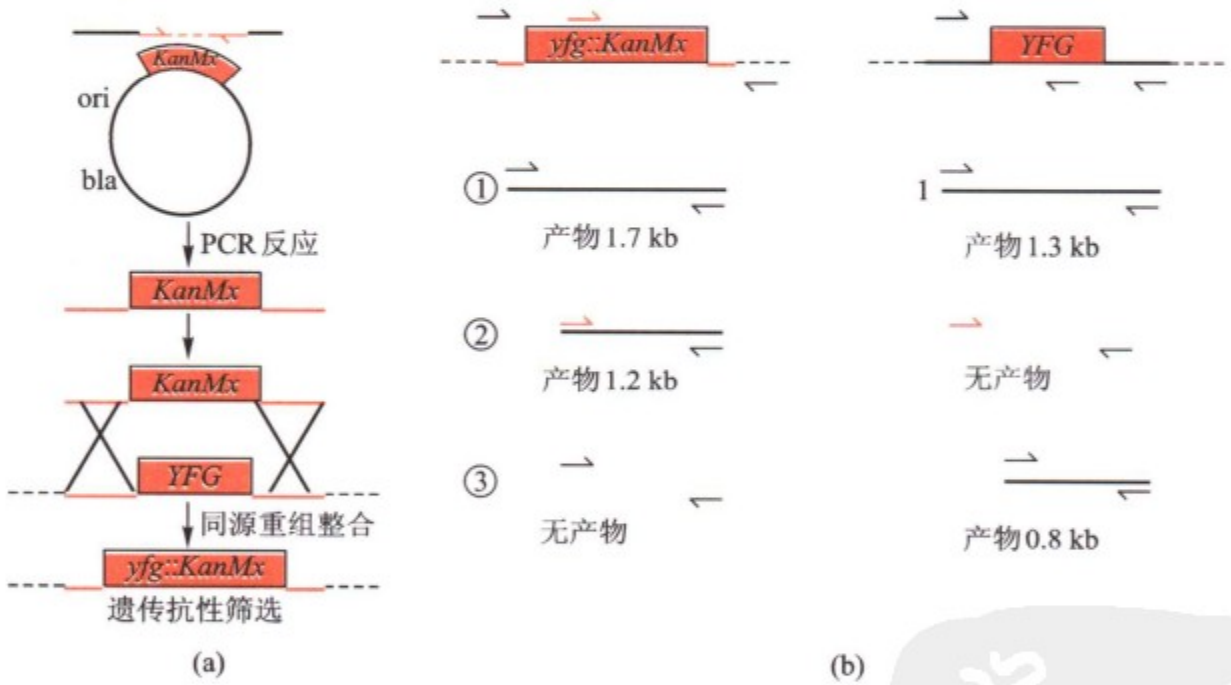



**图 6-32 "一步置换法"制备酵母缺失突变体的流程图** (a)基因置换。用 PCR 方法扩增 *KanMx* 基因片段(载体上该基因两侧为与目的基因 *YFG* 同源的序列),纯化 PCR 产物后转入酵母细胞,片段与目的基因发生同源重组,使 *KanMx* 基因片段插入酵母基因组,细胞具有 G418 抗性。(b)用 PCR 筛选基因置换事件。按如下三步验证是否在正确位置上发生重组:①采用目的基因两侧非翻译区的引物,PCR 得到不同大小的片段,有外源基因插入的 PCR 产物显著大于无外源基因插入时;②如用 *KanMx* 基因内部引物和目的基因一侧引物做 PCR,只能在发生正确重组的酵母细胞中得到 PCR 产物,没有发生重组的细胞中不能得到 PCR 产物;③如用目的基因被置换掉的那段部分序列为引物做 PCR,那么,发生重组的细胞中不能得到 PCR 产物,野生型细胞中有 PCR 产物。

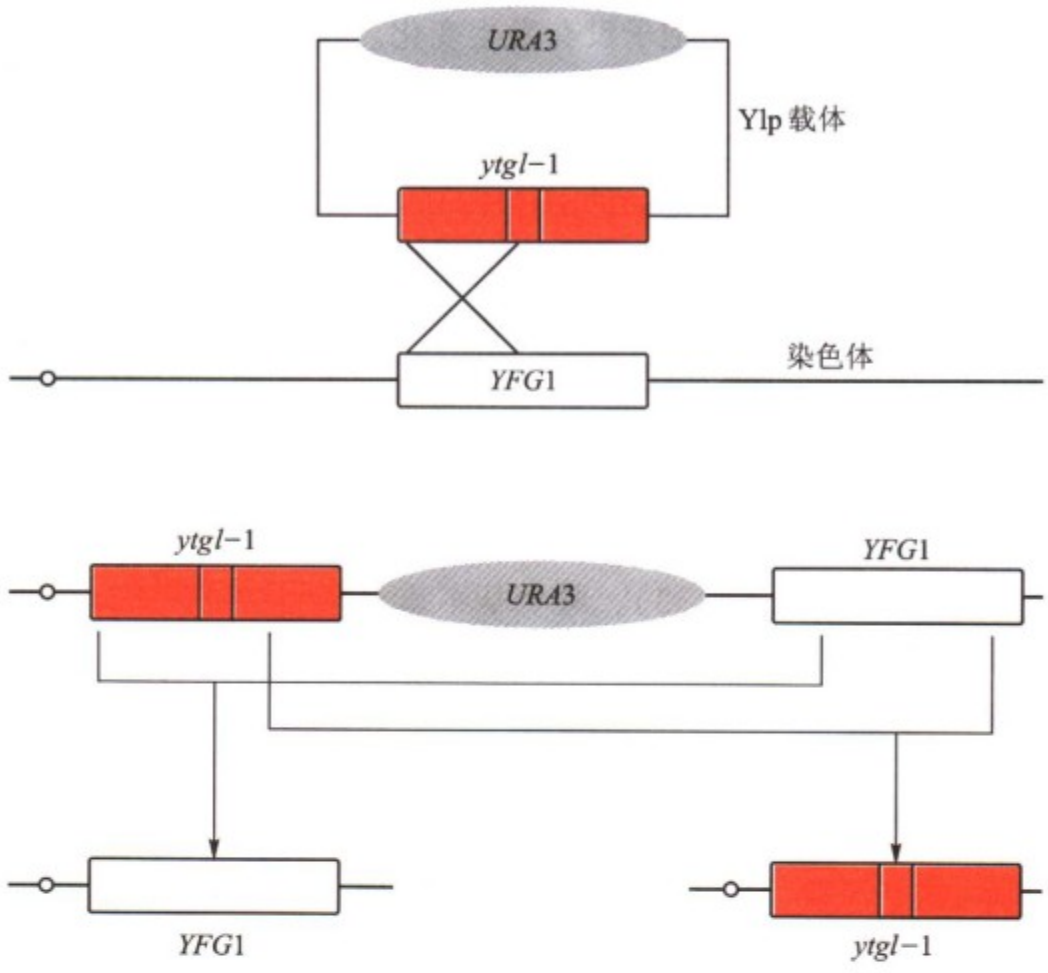



图 6-33 “两步置换法”制备酵母缺失突变体的流程图

为 URA3,用 5-氟乳清酸(5-FOA)培养基可挑选出丢失质粒 DNA 的菌株。细胞中如果只发生第一步重组反应,URA3 所产生的 XX 仍然会使 5-FOA 转变成对细胞有毒的物质,细胞被杀死。只有当细胞中连续发生了两步重组反应,URA3 被剔除,酵母细胞才能在5-FOA 培养基上生长。

### 6.5.3 外源基因在酵母中的功能鉴定

将外源基因克隆于酵母表达载体上,转化野生型或突变酵母菌株,通过观察酵母的表型变化或分析细胞中化学成分的变化即可推测该基因的生物学功能。例如,在酵母细胞中表达外源基因与绿色荧光蛋白基因的融合蛋白后,可通过荧光显微镜观察荧光所在的亚细胞区域,从而了解该基因产物在细胞中的定位。野生型酵母细胞中 18 碳不饱和脂肪酸主要是 C18:1,这些细胞不能产生含有两个或多个双键的不饱和脂肪酸,有人将棉花中可能编码 ω-6 脂肪酸脱氢酶的基因转入野生型酵母细胞中,诱导该基因的表达,然后用气相色谱和质谱技术分析细胞的脂肪酸组成发现,带有棉花基因的野生型酵母中出现了 C18:2 不饱和脂肪酸,证明该基因编码了有功能的 ω-6 脂肪酸脱氢酶(图 6-34)。因为酵母细胞中有与动植物细胞比较接近的蛋白质转录后修饰系统,许多在大肠杆菌中不产生功能蛋白质的动植物基因在酵母中表达后往往具有生物学功能(表 6-2)。

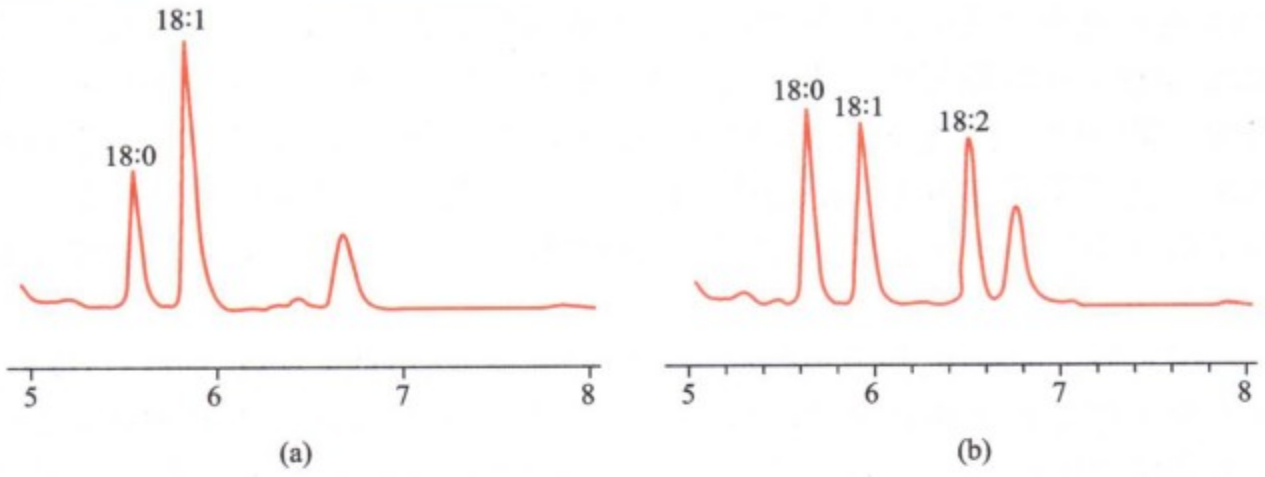



**图 6-34 用气相色谱法检测基因功能** (a) 将空载体转入无合成 C18:2 长链脂肪酸能力的酵母细胞后，用气相色谱证明细胞抽提物中只有 C18:1，没有 C18:2 不饱和脂肪酸。(b) 将 FAD2(fatty acid desaturase 2,为 ω-6 脂肪酸去饱和酶)载体转入酵母细胞后,C18:1 减少,出现大量 C18:2,说明该基因编码的蛋白可能以 C18:1 为底物合成 C18:2。

**表 6-2 大肠杆菌和酵母表达系统比较**

| 性　质 | 大肠杆菌 | 酵　母 |
|---|---|---|
| 细胞生长 | 快(30 min) | 快(90 min) |
| 培养基的复杂性 | 低 | 低 |
| 培养基的费用 | 低 | 低 |
| 蛋白表达水平 | 高 | 低 |
| 细胞外表达 | 分泌到细胞外周胞质中 | 分泌到培养基 |
| 蛋白折叠 | 经常需要再折叠 | 可能需要再折叠 |
| N-连接糖基化 | 无 | 在高甘露糖的条件下发生 |
| O-连接糖基化 | 无 | 有 |
| 磷酸化 | 无 | 有 |
| 乙酰化 | 无 | 有 |
| 酰基化 | 无 | 有 |
| γ-羧基化 | 无 | 有 |

## 6.6 其他分子生物学技术

### 6.6.1 凝胶滞缓实验

凝胶滞缓实验 (electrophoretic mobility shift assay,EMSA)，是体外分析 DNA 与蛋白质相互作用的一种特殊的凝胶电泳技术。在历史上，它首次为大肠杆菌乳糖阻遏物与其 DNA 结合位点的相互作用提供了动态分析数据。其基本原理是，蛋白质与 DNA 结合后将大大增加相对分子质量，而凝胶电泳中

DNA 朝正电极移动的距离与其相对分子质量的对数成正比。因此,没有结合蛋白的DNA 片段跑得快, 而与蛋白质形成复合物的 DNA 由于受到阻滞而跑得慢。实验中,当特定的 DNA 片段与细胞提取物混合后,若该复合物在凝胶电泳中的迁移速率变小,就说明该 DNA 可能与提取物中某个蛋白质分子发生了相互作用。该方法简单、快捷,是分离纯化特定 DNA 结合蛋白质的经典实验方法。

在 EMSA 试验中,用放射性同位素标记待检测的 DNA 片段(即探针DNA),然后与细胞提取物共温育,形成 DNA-蛋白质复合物,将该复合物加到非变性聚丙烯酰胺凝胶中进行电泳。通常用 $^{32}P$ 标记 DNA 分子,而不标记蛋白质。电泳结束后,用放射自显影技术显现具放射性标记的 DNA 条带位置。如果细胞蛋白提取物中不存与同放射性标记的 DNA 探针相结合的蛋白质, 那么所有放射性标记都将出现在凝胶的底部;反之,将会形成 DNA-蛋白质复合物,由于受到凝胶阻滞的缘故,放射性标记的 DNA 条带就会出现在凝胶的不同部位(图 6-35a)。

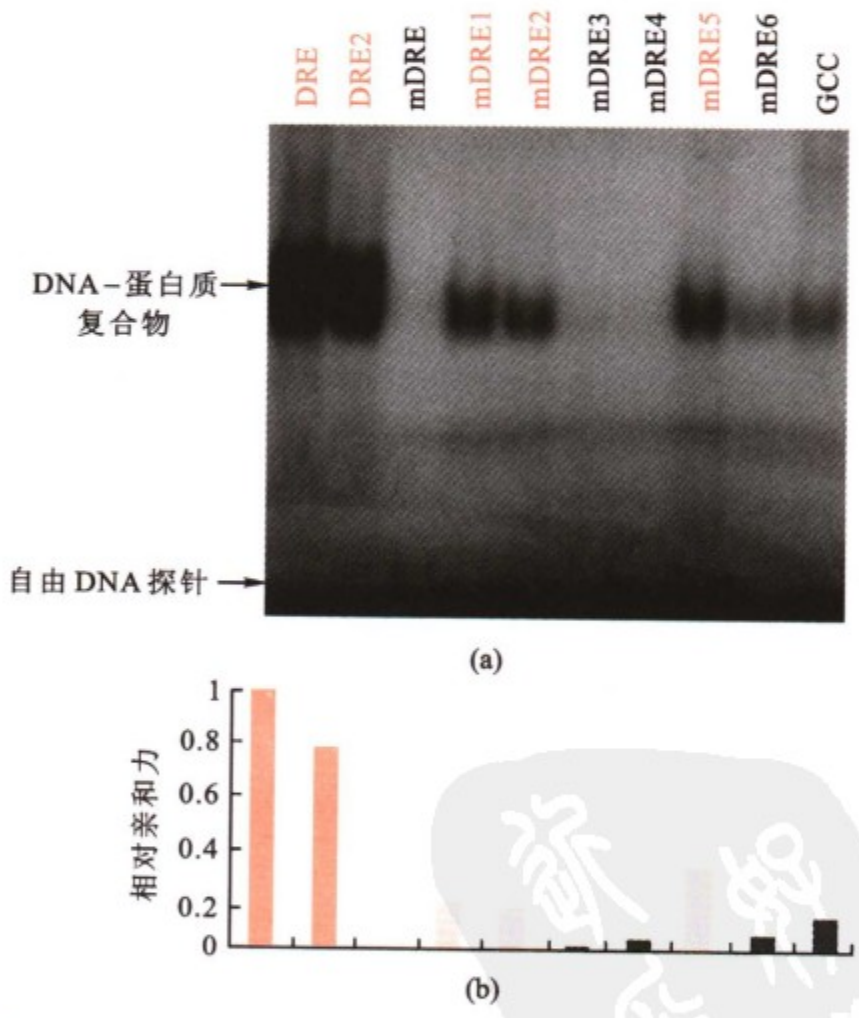



图 6-35 拟南芥转录因子 at5g11590 与不同 DNA 元件有不同的结合能力 (a) at5g11590 与不同 DNA 元件的EMSA 试验结果;(b) 灰度扫描分析显示 at5g11590 对不同 DNA 元件存在着很大的亲和力差异。

此外,EMSA 还被用于研究与蛋白质相结合的 DNA 序列的特异性。在 DNA-蛋白质复合物中加入超量的非标记竞争 DNA 分子, 如果它与标记的探

针DNA 结合同一种蛋白质,由于竞争 DNA(非标记)远远多于探针 DNA(标记),按照动力学原理,绝大部分蛋白质会与竞争 DNA 结合,使绝大部分探针 DNA 处于自由状态,凝胶电泳的放射自显影图上不会出现阻滞条带。相反,如果竞争DNA 不能与探针 DNA 所结合的蛋白质发生相互作用,那么,探针 DNA 将仍旧与其结合蛋白形成复合物,从而出现阻滞条带。另外,通过设计一系列引入一个或数个突变碱基的放射性标记探针,可以运用 EMSA 来评估这些突变对探针DNA 与结合蛋白相互作用的影响,进而确定该探针 DNA 分子中与蛋白质直接发生相互作用的关键性碱基(图 6-35b)。

## 6.6.2 噬菌体展示技术

噬菌体是细菌病毒的总称,英文为 bacteriophage,来源于希腊文“phagos”,有“吞噬”之意。作为细菌寄生物的噬菌体,它可以在脱离宿主细胞的状态下保持自己的生命。但是,一旦脱离了宿主细胞,它们就既不能生长也不能复制,因为大多数的噬菌体只能利用宿主核糖体、合成蛋白质的因子、各种氨基酸及能量代谢体系进行生长和增殖。噬菌体可被分为溶菌周期和溶源周期两种不同的类型。在溶菌周期,噬菌体将其感染的宿主细胞转变成为噬菌体的“制造厂”,产生出大量的子代噬菌体颗粒。实验室常将只具有溶菌生长周期的噬菌体叫做烈性噬菌体。而溶源生长周期是指噬菌体 DNA 被整合到宿主细胞染色体 DNA 上,成为它的一个组成部分,感染过程中不产生子代噬菌体颗粒。将具有这种溶源周期的噬菌体称为温和噬菌体。

噬菌体展示技术是将基因表达产物与亲和选择相结合的技术,其基本原理是将编码“诱饵”蛋白的 DNA 片段插入噬菌体基因组,并使之与噬菌体外壳蛋白编码基因相融合。该重组噬菌体侵染宿主细菌后,复制形成大量带有杂合外壳蛋白的噬菌体颗粒,直接用于捕获靶蛋白库中与“诱饵”相互作用的蛋白质(图 6-36)。噬菌体次要结构蛋白 pⅢ和主要结构蛋白(外壳蛋白)pⅧ均可作为载体来展示外源多肽。*p Ⅲ*基因融合时表达强度低(每个噬菌体上有不超过 5 拷贝),与 *p Ⅷ*基因融合时,表达强度高(每个噬菌体外壳上可能有 2 700~3 000 拷贝)。在 *p Ⅲ*和 *p Ⅷ*基因的信号肽与成熟蛋白质编码区之间都有单一的克隆位点便于外源基因插入,表达的融合蛋白被分泌到细胞间质并由宿主蛋白酶切除信号肽,融合蛋白被作为结构外壳蛋白呈现到病毒粒子的表面。

亲和筛选包括直接法和间接法。直接法是将靶蛋白质分子偶联到固相支持物上,文库噬菌体与固相支持物温育后洗去未结合的噬菌体,即获得与所筛选蛋白有亲和性的噬菌体。间接法是将生物素标记的蛋白质分子与文库噬菌体温育后铺在含有链亲和素(能与生物素相结合)的平皿上,洗去未结合的噬菌体,保留在平皿上的就是结合状态的噬菌体。洗脱结合状态噬菌体后感染细菌,扩增噬菌体,开始新一轮的筛选,连续几次亲和纯化反应后就能选择性富集并扩增结合靶蛋白的噬菌体。

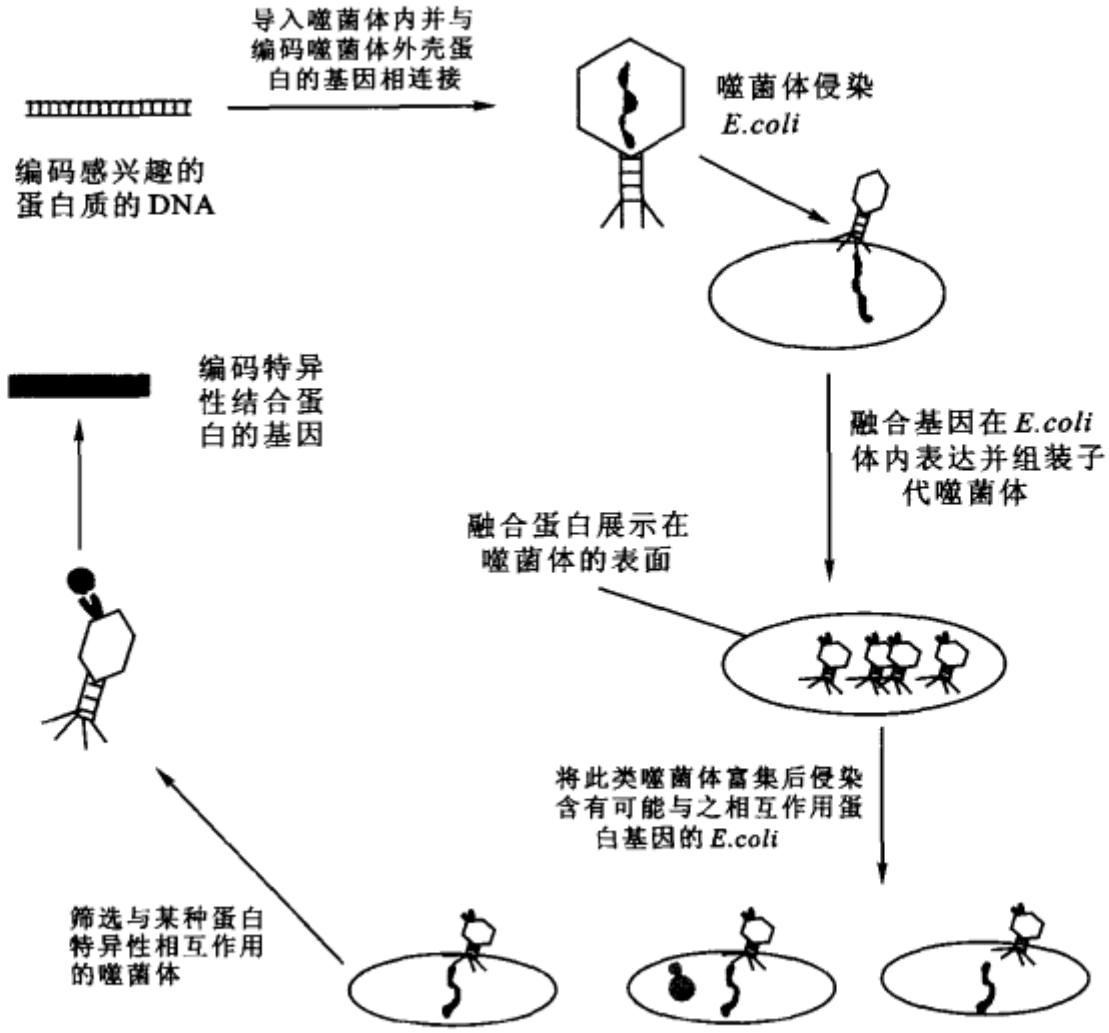



图 6-36 噬菌体展示技术研究蛋白质相互作用

## 6.6.3 蛋白质磷酸化分析技术

由蛋白激酶和蛋白磷酸酯酶所催化的蛋白质可逆磷酸化过程,是生物体内一种普遍且重要的调节方式，控制着众多的生理生物化学反应和生物学过程。细胞的生长发育、周期调控、基因表达、蛋白合成以及神经功能、肌肉收缩等都离不开蛋白质特异性磷酸化。许多复杂的细胞信号转导系统还涉及由多个蛋白激酶所催化的磷酸化级联反应。通常蛋白激酶催化 ATP 或 GTP 的 γ-磷酸基团转移到底物蛋白的丝氨酸、苏氨酸或酪氨酸残基上,促使底物蛋白发生磷酸化,而蛋白质磷酸酯酶则能催化底物蛋白发生去磷酸化(图 6-37)。此外,底物蛋白的组氨酸、赖氨酸、精氨酸、天冬氨酸、谷氨酸和半胱氨酸残基上也能发生可逆磷酸化。

实验室研究蛋白质磷酸化主要包括底物蛋白磷酸化和蛋白激酶活性检测,一般采用双向电泳及质谱分析等检测底物蛋白磷酸化。由于细胞内可能有多达上千个蛋白激酶,所以体内检测某个蛋白激酶活性非常困难,科学家常用体外激酶活性分析法来检测蛋白激酶活性。该方法能十分可靠地鉴定某个蛋白激酶对底物的磷酸化作用,筛查激酶的特异性底物。蛋白激酶体外磷酸化活性分析如图 6-38 所示，主要分为获取底物蛋白或待检测蛋白激酶和进行蛋白质体外磷酸化反应两步。目的蛋白(包括底物蛋白及待检测蛋白激酶)可以直接从组织

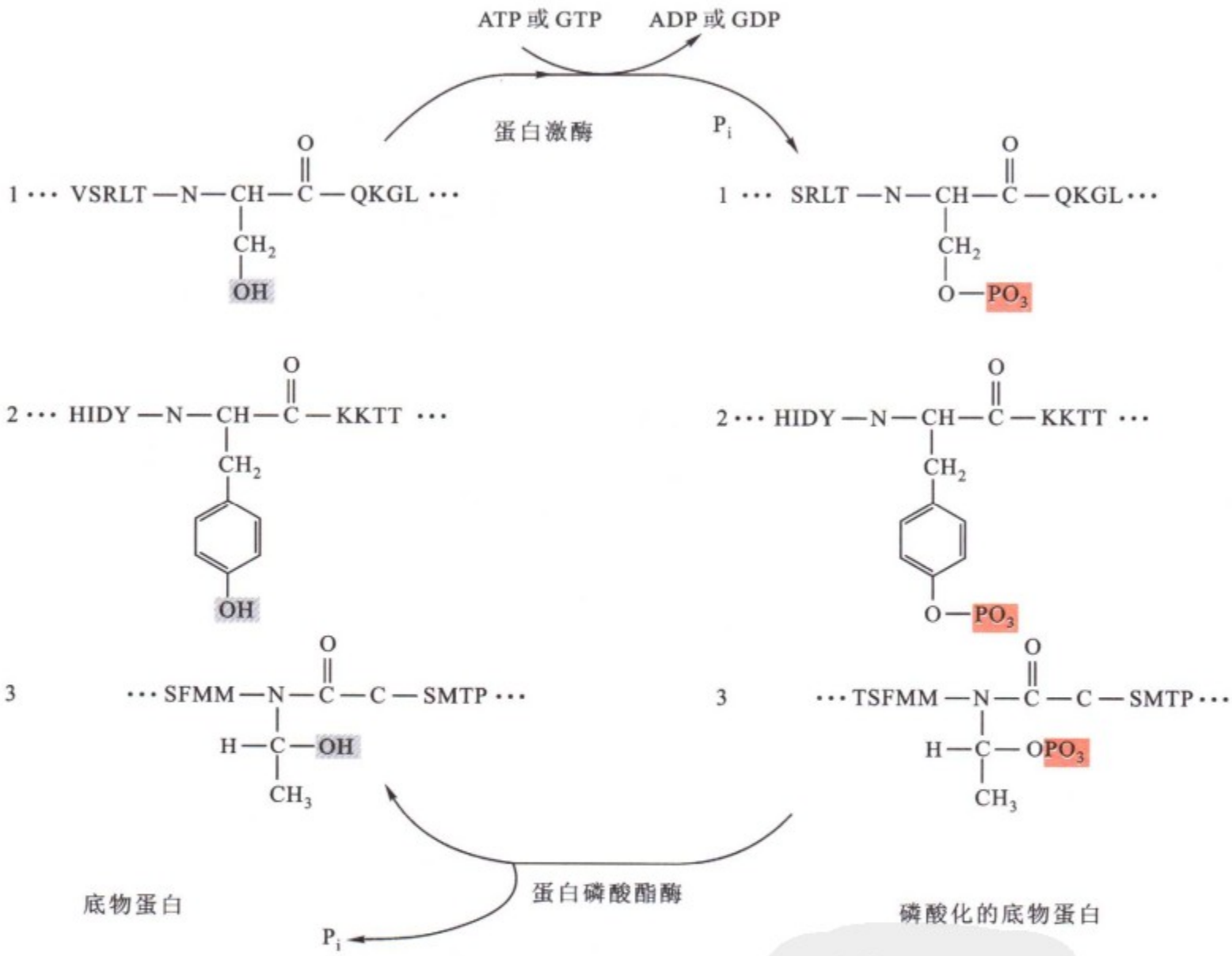
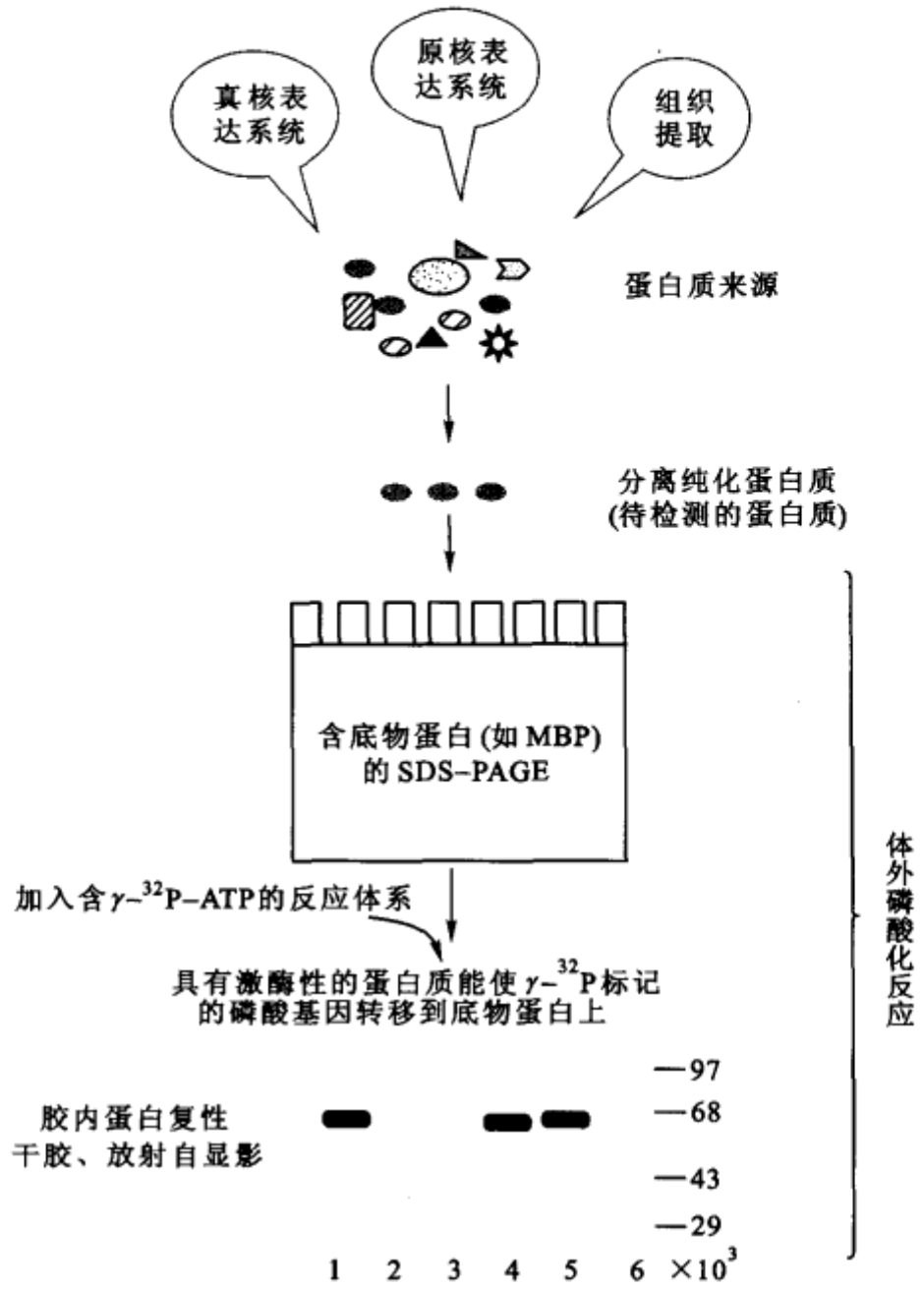



**图 6-37 蛋白激酶和蛋白磷酸酯酶催化底物蛋白发生可逆磷酸化** 蛋白质 1 的 Ser 残基上发生可逆磷酸化；蛋白质 2 的 Tyr 残基上发生可逆磷酸化；蛋白质 3 的 Thr 残基上发生可逆磷酸化。




▌图 6-38 蛋白激酶体外磷酸化活性分析 泳道 1、4 和 5 中的蛋白质具有使蛋白质磷酸化的激酶活性，泳道 2、3 和 6 中的蛋白质不能使 MBP(myelin basic protein)磷酸化。

中提取，也可由原核或真核表达载体诱导表达得到。

## 思考题

1. 列举两种研究基因表达模式的方法并简述其原理。
2. 原位杂交技术中哪些步骤可以降低探针与 RNA 的非特异性结合？
3. 简述基因芯片技术对分子生物学研究的意义。
4. 比较酵母双杂交技术和免疫共沉淀技术在研究蛋白质相互作用方面的优缺点。
5. 简述 RNAi 技术在分子生物学领域的应用前景和存在的问题。
6. 假设某哺乳动物基因可能编码了脂肪酸脱氢酶，请在酵母细胞中设计试验

研究该基因的功能。

7. 列举三种研究 DNA 和蛋白质相互作用的方法，并比较其优缺点。

8. 一个二倍体酵母细胞存在两对不相关的等位基因 *Aa* 和 *Bb*（大写字母代表野生型酵母基因，小写字母代表突变基因），请列出该二倍体细胞可能产生哪些不同基因型的四分体细胞？

9. 列举研究蛋白质磷酸化的主要实验方法并比较其优缺点。

## 主要参考文献


1. Fernandez-Gacio A, Uguen M, Fastrez J. Phage display as a tool for the directed evolution of enzymes. Trends in Biotechnology, 2003, 21:408–414.
2. Burke D, Dawson D, Stearns T. Methods in yeast genetics. New York: Cold Spring Harbor Press, 2000.
3. No D, Yao T P, Evans R M. Ecdysone-inducible gene expression in mammalian cells and transgenic mice. Proc. Nat. Acad. Sci. U. S. A. 1996, 93:3346–3351.
4. Elbashir S M, Lendeckel W, Tuschl T. RNA interference is mediated by 21 and 22 nt RNAs. Genes Dev. 2001, 15:188–200.
5. Fire A, Xu S, Montgomery M K, *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature, 1998, 391:806–811.
6. Montgomery M K. RNA interference: unraveling a mystery. Nat. Struct. Mol. Biol. 2006, 13:1039–1041.

# 7 基因的表达与调控(上)

## ——原核基因表达调控模式


- 7.1 原核基因表达调控总论
- 7.2 乳糖操纵子与负控诱导系统
- 7.3 色氨酸操纵子与负控阻遏系统
- 7.4 其他操纵子
- 7.5 固氮基因调控
- 7.6 转录水平上的其他调控方式
- 7.7 转录后调控


原核生物及单细胞真核生物直接暴露在变幻莫测的环境中，食物供应无保障，只有能根据环境条件的改变合成各种不同的蛋白质，使代谢过程适应环境的变化，才能维持自身的生存和繁衍。高等真核生物代谢途径和食物来源都比较稳定，但由于它们是多细胞有机体，在个体发育过程中出现细胞分化，形成各种组织和器官，而不同类型的细胞所合成的蛋白质在质和量上都是不同的。因此，不论是真核还是原核细胞都有一套准确地调节基因表达和蛋白质合成的机制。

自然选择倾向于保留高效率的生命过程。在单细胞生物中，使细胞代谢总效率增加的突变细胞生长略快于野生型，只要有足够的时间，突变细胞系将在整个细胞群体占主导地位。例如，在一个每 30 min 增殖一倍的 $10^9$ 细菌群体中，若有一个细菌变成了 29.5 min 增殖一倍，大约经过 80 d 的连续生长后，这个群体中的 99.9%都将具有 29.5 min 增殖一倍的生长速度（假设培养液被不断地补充，否则细菌将在一两天内停止生长）。

原核生物细胞的基因和蛋白质种类较少，如大肠杆菌基因组约为 $4.60\times10^6$ bp 共有 4 288 个可译框架。据估计，一个原核细胞中总共含有 $10^7$ 个蛋白质分子，如果每个基因等同翻译的话，任何一个多肽应有 2 500 拷贝。但是，这些蛋白质并不是以相同拷贝数存在于每个细胞中的，有些蛋白质的数目相当固定，另一些则变化很大。例如，每个大肠杆菌细胞可以有约 15 000 个核糖体，与其结合的约 50 种核糖体蛋白数量也是十分稳定的。糖酵解体系的酶含量很大，其数目也极恒定。此外，如 DNA 聚合酶、RNA 聚合酶等都是代谢过程中十分必需的酶或蛋白质，其合成速率不

受环境变化或代谢状态的影响，这一类蛋白质被称为永久型(constitutive)合成蛋白质。另一种类型则被称为适应型或调节型(adaptive or regulated)，因为这类蛋白质的合成速率明显地受环境的影响。例如，一般情况下，一个大肠杆菌细胞中只有 15 个分子的β-半乳糖苷酶，但若将细胞培养在只含乳糖的培养基中，每细胞中这种酶的量可高达几万个分子。其他参与糖代谢的酶，氨基酸、核苷酸合成系统的酶类，其合成速度和总量都随着培养条件的变化而改变。

细菌中所利用的大多数基本调控机制一般执行如下规律：一个体系在需要时被打开，不需要时被关闭。这种“开-关”(on-off)活性是通过调节转录来建立的，也就是说 mRNA 的合成是可以被调节的。实际上，当我们说一个系统处于“off”状态时，也可能有本底水平的基因表达，常常是每世代每个细胞只合成 1 或 2 个 mRNA 分子和极少量的蛋白质分子。为了方便，我们常常使用“off”这一术语，但必须明白所谓“关”实际的意思是基因表达量特别低，很难甚至无法检测到。

## 7.1 原核基因表达调控总论

我们知道，所有生物的遗传信息，都是以基因的形式储藏在细胞内的 DNA (或 RNA) 分子中。随着个体的发育，DNA 分子能有序地将其所承载的遗传信息，通过密码子-反密码子系统，转变成蛋白质或功能 RNA 分子，执行各种生理生物化学功能，完成生命的全过程。科学家把这个从 DNA 到蛋白质或功能 RNA 的过程称为基因表达(gene expression，图 7-1)，对这个过程的调节就称为基因表达调控(gene regulation 或 gene control)。基因调控是现阶段分子生物学研究的中心课题。要了解动、植物生长发育的规律、形态结构特征和生物学功能，就必须弄清楚基因表达调控的时间和空间概念。

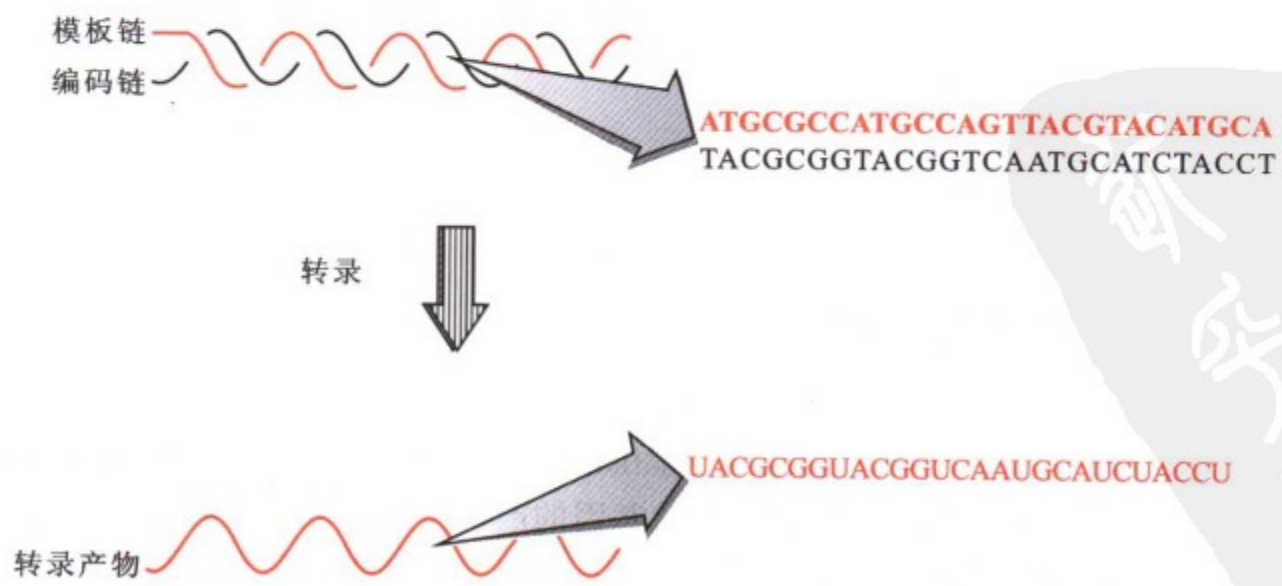



图 7-1 基因表达的第一步是由 RNA 聚合酶拷贝 DNA 双链中的模板链，生成与该序列完全互补(除了 T 被置换成 U 之外)的 RNA 链

基因表达调控主要表现在以下两个方面：

1. 转录水平上的调控(transcriptional regulation)。

2. 转录后水平上的调控(post-transcriptional regulation)，包括：

① mRNA 加工成熟水平上的调控(differential processing of RNA transcript)。

② 翻译水平上的调控(differential translation of mRNA)。

基因调控的指挥系统也是多样的，不同的生物使用不同的信号来指挥基因调控。原核生物中，营养状况(nutritional status)和环境因素(environmental factor)对基因表达起着举足轻重的影响。在真核生物尤其是高等真核生物中，激素水平(hormone level)和发育阶段(developmental stage)是基因表达调控的最主要手段，营养和环境因素的影响力大为下降。

在转录水平上对基因表达的调控决定于 DNA 的结构、RNA 聚合酶的功能、蛋白因子及其他小分子配基的相互作用。在研究转录及转录调控以前，我们先来分析一下原核与真核生物转录与翻译的特点。因为细菌 mRNA 在形成过程中与核糖体混合在一起，所以，细菌的转录与翻译过程几乎发生在同一时间间隔内，转录与翻译相偶联 (coupled transcription and translation)。真核生物中，转录产物(primary transcript)只有从核内运转到核外，才能被核糖体翻译成蛋白质(图 7-2)。

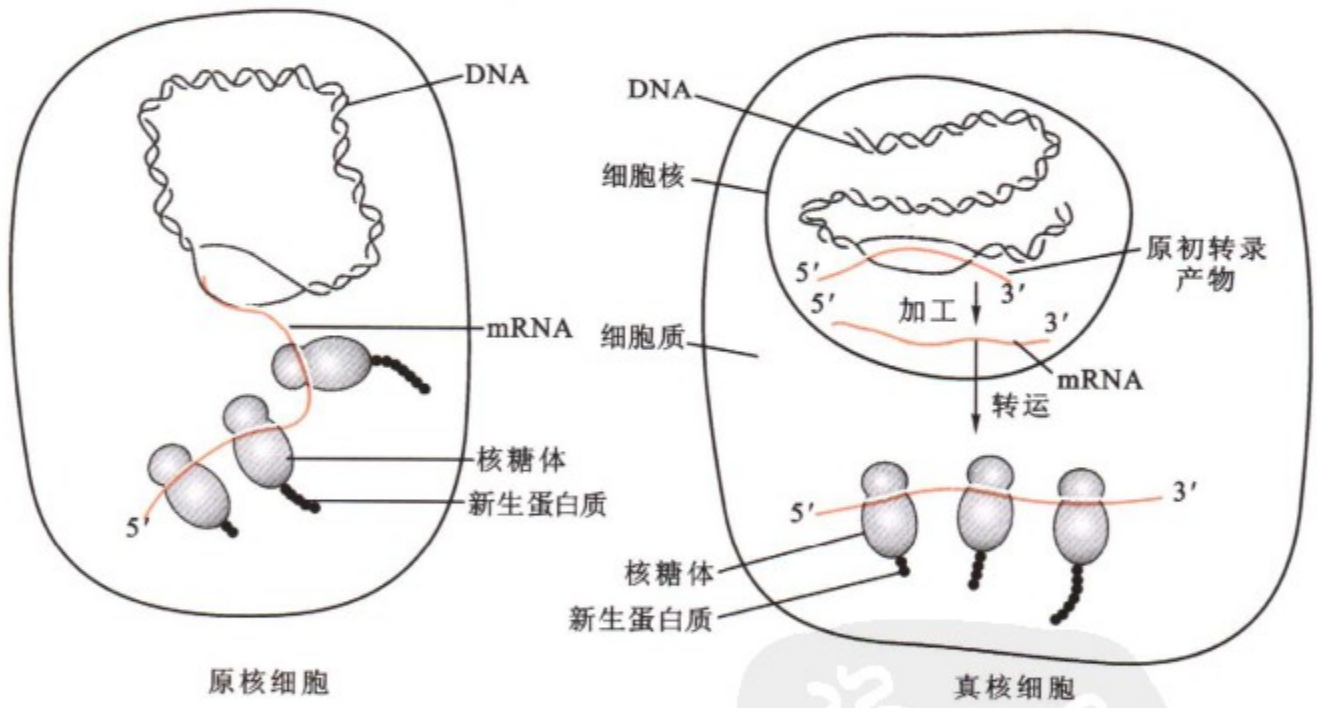



图 7-2 真核与原核生物转录及翻译调控的总体特征比较

## 7.1.1 原核基因调控分类

原核生物的基因调控主要发生在转录水平上，根据调控机制的不同可分为负转录调控(negative transcription regulation)和正转录调控(positive transcription regulation)。在负转录调控系统中，调节基因的产物是阻遏蛋白(repressor)，起着阻止结构基因转录的作用。根据其作用特征又可分为负控诱导和负控阻遏二大类(表 7-1)。在负控诱导系统中，阻遏蛋白不与效应物(诱导物)结合时，结构基因不转录；在负控阻遏系统中，阻遏蛋白与效应物结合时，结构基因不转录。阻

表 7-1 负转录调控的类型及其特点

| 调节物结合到 DNA 上 | 正调节 | 负调节 |
| --- | --- | --- |
| 是 | 开启 | 关闭 |
| 否 | 关闭 | 开启 |

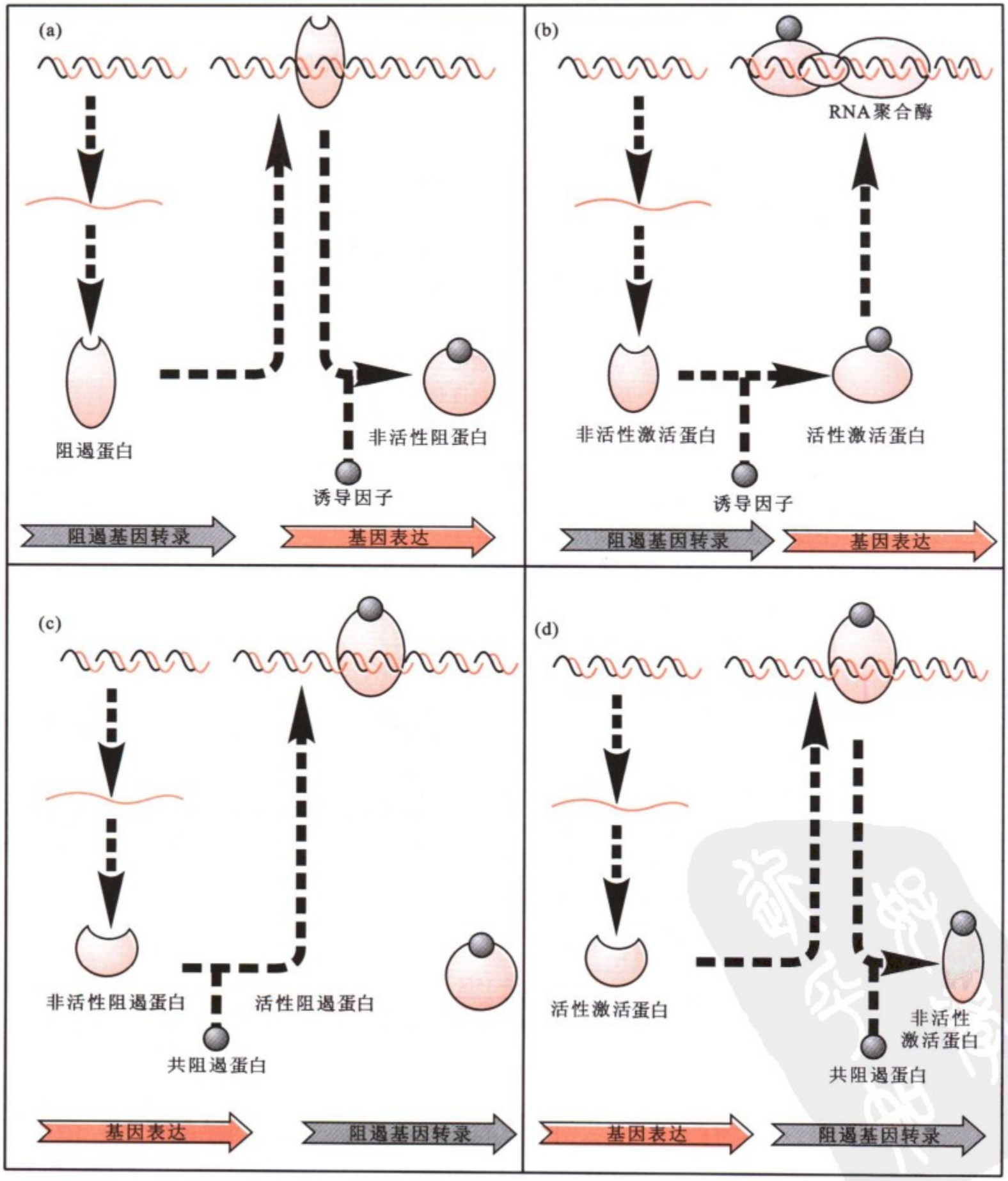



图 7-3 细菌中的转录调控体系 (a)负控诱导系统;(b)正控诱导系统;(c)负控阻遏系统;(d)正控阻遏系统。

遏蛋白作用的部位是操纵区。在正转录调控系统中,调节基因的产物是激活蛋白(activator)。也可根据激活蛋白的作用性质分为正控诱导系统和正控阻遏系统。在正控诱导系统中,效应物分子(诱导物)的存在使激活蛋白处于活性状态;在正控阻遏系统中,效应物分子的存在使激活蛋白处于非活性状态(图 7-3)。

参与大肠杆菌中基因表达调控最常见的蛋白质可能是 $\sigma$ 因子,基因组序列分析后发现至少存在六种 $\sigma$ 因子,并根据其相对分子质量的大小或编码基因进行命名(表 7-2),其中 $\sigma^{70}$ 参与最基本的生理功能如碳代谢、生物合成等基因的转录调控。

表 7-2 大肠杆菌中的各种 $\sigma$ 因子比较

| $\sigma$ 因子 | 编码基因 | 主 要 功 能 |
|---|---|---|
| $\sigma^{70}$ | *RpoD* | 参与对数生长期和大多数碳代谢过程基因的调控 |
| $\sigma^{54}$ | *RpoN* | 参与多数氮源利用基因的调控 |
| $\sigma^{38}$ | *RpoH* | 分裂间期特异基因的表达调控 |
| $\sigma^{32}$ | *RpoS* | 热休克基因的表达调控 |
| $\sigma^{28}$ | *RpoF* | 鞭毛趋化相关基因的表达调控 |
| $\sigma^{24}$ | *RpoE* | 过度热休克基因的表达调控 |

除参与氮代谢的 $\sigma^{54}$ 以外,其他 5 种 $\sigma$ 因子在结构上具有同源性,所以统称 $\sigma^{70}$ 家族。研究发现,所有 $\sigma$ 因子都含有 4 个保守区(图 7-4),其中第 2 个和第 4 个保守区参与结合启动区 DNA,第 2 个保守区的另一部分还参与双链 DNA 解开成单链的过程。与上述 $\sigma$ 因子特异性结合 DNA 上的-35 区和-10 区不同,$\sigma^{54}$ 因子识别并与 DNA 上的-24 和-12 区相结合(图 7-5)。在与启动子结合的顺序上,$\sigma^{70}$ 类启动子在核心酶结合到 DNA 链上之后才能与启动子区相结合,而 $\sigma^{54}$ 则类似于真核生物的 TATA 区结合蛋白(TBP),可以在无核心酶时独立结合到启动子上。

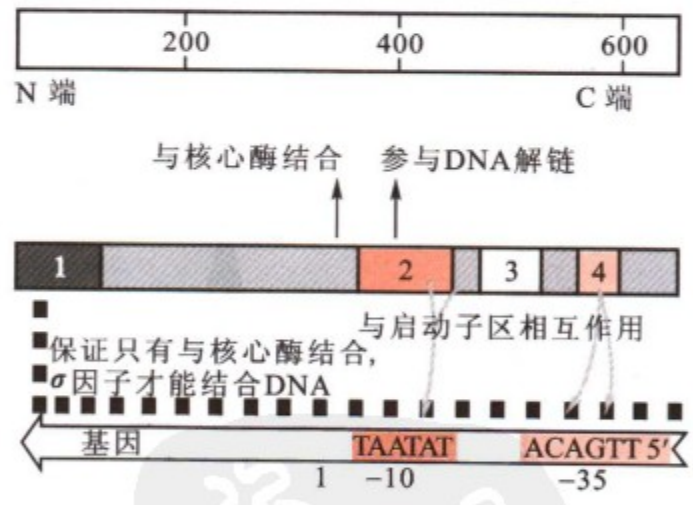



图 7-4 大肠杆菌中的 $\sigma$ 因子(以 $\sigma^{70}$ 为例)主要包括 4 个结构域

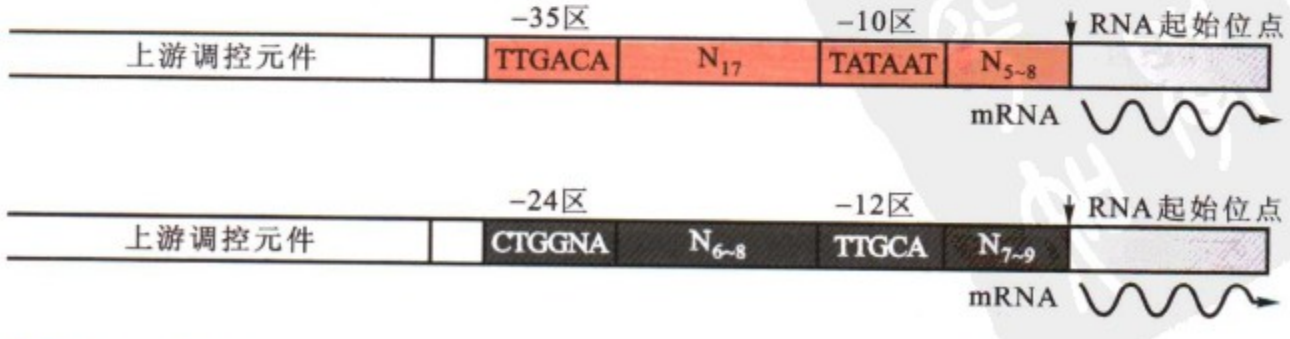



图 7-5 $\sigma^{70}$(上)和 $\sigma^{54}$(下)因子识别并结合在所调控基因上游的不同区域

### 7.1.2 原核基因调控的主要特点

原核生物通过特殊代谢物调节基因活性主要分为可诱导调节和可阻遏调节：

1. 可诱导调节

是指一些基因在特殊的代谢物或化合物的作用下，由原来关闭的状态转变为工作状态，即在某些物质的诱导下使基因活化。这类基因中最突出的例子是大肠杆菌的乳糖操纵子。大肠杆菌在含有葡萄糖的培养基中生长良好，在只含乳糖的培养基中开始时生长不好，直到合成了能够利用乳糖的一系列酶，具备了利用乳糖作为碳源的能力才开始生长。细菌获得这一能力的原因就是因为在诱导物乳糖的诱导下启动了乳糖操纵子，表达它所编码的 3 个酶：β-半乳糖苷酶(使乳糖水解为半乳糖和葡萄糖)，β-半乳糖苷透过酶(使乳糖进入细菌细胞内)，β-半乳糖苷乙酰基转移酶(使β-半乳糖第六位碳原子乙酰化)。这个操纵子启动的效果是十分明显的。以β-半乳糖苷酶为例，在葡萄糖培养基中生长时，每个细胞只有几个酶分子，但若转移到乳糖培养基中，几分钟后，每个细菌细胞内可产生 3 000 个酶分子。因此，这类基因被称为可诱导基因，这类酶被称为诱导酶，这个生物化学过程被称为酶的诱导合成。

2. 可阻遏调节

这类基因平时都是开启的，处在产生蛋白质或酶的工作过程中，由于一些特殊代谢物或化合物的积累而将其关闭，阻遏了基因的表达。比如大肠杆菌生活中必须有色氨酸，一般情况下，该操纵子是开启的。如果在细菌培养基中加入色氨酸，使之能利用培养基中的色氨酸来维持生活而不需要再费力去合成，细菌往往能在 2~3 min 内完全关闭该操纵子。这就是说，某一代谢途径最终产物合成酶的基因可以被这个产物本身所关闭。这种基因被称为可阻遏基因，这些酶被称为可阻遏酶，这个现象被称为可阻遏现象，这些起阻遏作用的小分子被称为阻遏物。在这里，色氨酸或与其代谢有关的某种物质在阻遏过程(而不是诱导过程)中起作用。

作为一般规律，可诱导的操纵子总是一些编码糖和氨基酸分解代谢蛋白的基因，这些糖和氨基酸平时含量很少，细菌总是利用更普遍的能源物质——葡萄糖的水解来提供能源，因此，这些操纵子常常是关闭的。一旦生存条件发生变化，如葡萄糖缺乏而必须利用乳糖作为能源时，就要打开这些基因。可阻遏基因的情况恰好相反，它们是一些合成各种细胞代谢过程中所必需的小分子物质(如氨基酸、嘌呤和嘧啶等)的基因，由于这类物质在生命过程中的重要地位，这些基因总是打开着的。只有当细菌生活环境中有充分供应时，才关闭这些基因，停止其合成。

### 7.1.3 弱化子对基因活性的影响

属于这种调节方式的有大肠杆菌中的色氨酸操纵子、苯丙氨酸操纵子、苏氨酸操纵子、异亮氨酸操纵子和缬氨酸操纵子以及沙门菌的组氨酸操纵子和亮

氨酸操纵子、嘧啶合成操纵子等。

在这种调节方式中，起信号作用的是有特殊负载的氨基酰-tRNA 的浓度，在色氨酸操纵子中就是色氨酰-tRNA 的浓度。当操纵子被阻遏，RNA 合成被终止时，起终止转录信号作用的那一段核苷酸被称为弱化子。这种因为核糖体在基因转录产物上的不同位置，决定了 RNA 可以形成哪一种形式的二级结构，并由此决定基因能否继续转录的调节方式确实是非常巧妙的。起调节作用的信号分子是细胞中某一种氨基酸或嘧啶的浓度，因此是转录调节中的微调整，好比无线电调谐器中的微调一样，只要稍加变动就可影响整个体系的功能。我们将在后面分析色氨酸基因操纵子结构时加以详述。

### 7.1.4 降解物对基因活性的调节

操纵子学说的核心是使基因从表达抑制状态中解脱出来进行转录，是从负调节的角度来考虑基因表达调控的。那么，有没有从相反的角度进行正调节以提高基因的转录水平，使它由原来的低水平表达变成高水平表达的呢？这就是降解物抑制作用的调节。

葡萄糖存在的情况下，即使在细菌培养基中加入乳糖、半乳糖、阿拉伯糖或麦芽糖等诱导物，与其相对应的操纵子也不会启动，不会产生出代谢这些糖的酶来，这种现象称为葡萄糖效应或降解物抑制作用。为什么会产生这种效应呢？因为添加葡萄糖后，细菌所需要的能量便可从葡萄糖得到满足，葡萄糖是最方便的能源，细菌无需开启一些不常用的基因去利用这些稀有的糖类。葡萄糖的存在会抑制细菌的腺苷酸环化酶活性，减少环腺苷酸(cAMP)的合成，与它相结合的蛋白质——环腺苷酸受体蛋白(CRP)又称分解代谢物激活蛋白(CAP)，因找不到配体而不能形成复合物。降解物抑制作用是通过提高转录强度来调节基因表达的，是一种积极的调节方式。

### 7.1.5 细菌的应急反应

以上均是细菌处于正常生活条件下的基因表达调节方式。所谓“正常”，也包括生活环境中缺少某一种或两种能源，但能找到其他代用物的情况。可是，细菌有时会碰到紧急状况，比如氨基酸饥饿时，就不是缺少一二种氨基酸，而是氨基酸的全面匮乏。为了紧缩开支，渡过难关，细菌将会产生一个应急反应，包括生产各种 RNA、糖、脂肪和蛋白质在内的几乎全部生物化学反应过程均被停止。实施这一应急反应的信号是鸟苷四磷酸(ppGpp)和鸟苷五磷酸(pppGpp)。产生这两种物质的诱导物是空载 tRNA。

当氨基酸饥饿时，细胞中便存在大量的不带氨基酸的 tRNA，这种空载的 tRNA 会激活焦磷酸转移酶，使 ppGpp 大量合成，其浓度可增加 10 倍以上。ppGpp 的出现会关闭许多基因，当然也会打开一些合成氨基酸的基因，以应付这种紧急状况。关于 ppGpp 的作用原理还不大清楚，一般认为 RNA 聚合酶有不同的构型，这些构象可以识别不同的启动子区，ppGpp 与 RNA 聚合酶的结合会使它的某些构象稳定下来，从而改变了基因转录的效率。也有人认为基因转录

起始位点附近有一些保守序列，它们可能是 ppGpp 或有关调节蛋白的结合位点。当这些序列与 ppGpp 结合后，就不能再与 RNA 聚合酶结合，使基因被关闭。ppGpp 与pppGpp 的作用范围十分广泛，它们不是只影响一个或几个操纵子，而是影响一大批操纵子，所以它们是超级调控因子。

## 7.2 乳糖操纵子与负控诱导系统

操纵子学说是关于原核生物基因结构及其表达调控的学说，由法国巴斯德研究所著名科学家 Jacob 和 Monod 在 1961 年首先提出，并在 10 年内经许多科学家的补充和修正得以逐步完善。我们将在本节介绍建立这一模型的研究过程及实验依据，从而帮助读者了解原核生物基因表达调控的主要模式和机制。

大肠杆菌乳糖操纵子(lactose operon)包括 3 个结构基因：*Z*、*Y* 和 *A*，以及启动子、控制子和阻遏子等，如图 7-6 所示。转录时，RNA 聚合酶首先与启动区(promoter，P)结合，通过操纵区(operator，O)向右转录。转录从 O 区中间开始，按 *Z*→*Y*→*A* 方向进行，每次转录出来的一条 mRNA 上都带有这 3 个基因。转录的调控是在启动区和操纵区进行的。

| | P | *lacI* | *PO* | *lacZ* | *lacY* | *lacA* |
|---|---|---|---|---|---|---|
| DNA | | 1 040 | 82 | 3 510 | 780 | 825 |
| mRNA | | | | | | |
| 多肽 | | $3.8\times10^4$ | | $1.25\times10^5$ | $3.0\times10^4$ | $3.0\times10^4$ |
| 蛋白质 | | 四聚体 $5.0\times10^5$ | | 四聚体 $5.0\times10^5$ | 膜蛋白 $3.0\times10^4$ | 二聚体 $6.0\times10^4$ |
| 功能 | | 阻遏子 | | β-半乳糖苷酶 | 透过酶 | 乙酰基转移酶 |

图 7-6 *lac* 操纵子主要组分分析

3 个结构基因各决定一种酶：*Z* 编码 β-半乳糖苷酶；*Y* 编码 β-半乳糖苷透过酶；*A* 编码 β-半乳糖苷乙酰基转移酶。

β-半乳糖苷酶是一种 β-半乳糖苷键的专一性酶，除能将乳糖水解成葡萄糖和半乳糖外，还能水解其他 β-半乳糖苷(如苯基半乳糖苷)。β-半乳糖苷透过酶的作用是使外界的 β-半乳糖苷(如乳糖)透过大肠杆菌细胞壁和原生质膜进入细胞内，所以，如果用乳糖为大肠杆菌生长的唯一碳源和能源，这两种酶是必需的。β-半乳糖苷乙酰基转移酶的作用是把乙酰辅酶 A 上的乙酰基转移到 β-半乳糖苷上，形成乙酰半乳糖，它在乳糖的利用中并非必需。

### 7.2.1 酶的诱导——*lac* 体系受调控的证据

在不含乳糖及β-半乳糖苷的培养基中，*lac*⁺基因型大肠杆菌细胞内β-半乳糖苷酶和β-半乳糖苷透过酶的浓度很低，每个细胞内大约只有 1~2 个酶分子。但是，如果在培养基中加入乳糖，酶的浓度很快达到细胞总蛋白量的 6%或7%，每个细胞中可有超过 $10^5$ 个酶分子。

当有乳糖供应时，在无葡萄糖培养基中生长的 *lac*⁺细菌中将同时合成β-半乳糖苷酶和β-半乳糖苷透过酶，如图 7-7 所示。进一步用 $^{32}P$ 标记的 mRNA 与模板 DNA 进行定量分子杂交，表明培养基中加入乳糖 1~2 min 后，编码β-半乳糖苷酶和透过酶的 *lac* mRNA 量就迅速增加。去掉乳糖后，*lac* mRNA 量立即下降到几乎无法检测到，表明乳糖确实能激发 *lac* mRNA 的合成。

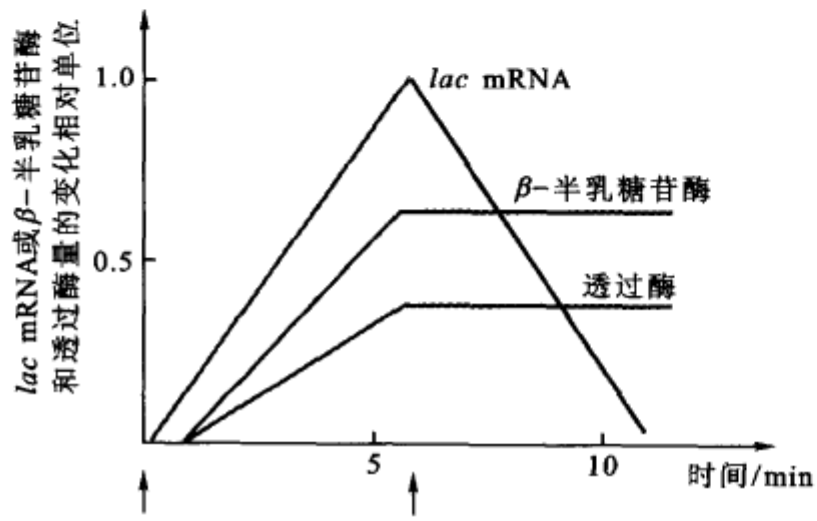



图 7-7 *lac* 体系的"开-关"特性

加入乳糖 1 min 后体系内就出现 *lac* mRNA，稍后即产生β-半乳糖苷酶和β-半乳糖苷透过酶。当移去乳糖之后，*lac* mRNA 总量立即下降，两种酶的活性却由于蛋白质半衰期较长而在相当长时期内维持恒定。

事实上，研究诱导作用时很少使用乳糖，因为培养基中的乳糖会被诱导合成的β-半乳糖苷酶所催化降解，从而使其浓度不断发生变化。实验室里常常使用两种含硫的乳糖类似物——异丙基巯基半乳糖苷(IPTG)和巯甲基半乳糖苷(TMG)。另外，在酶活性分析中，常用显色底物 *O*-硝基半乳糖苷(ONPG)。研究发现，它们都是高效诱导物(图 7-8)，因为它们都不是半乳糖苷酶的底物，所以

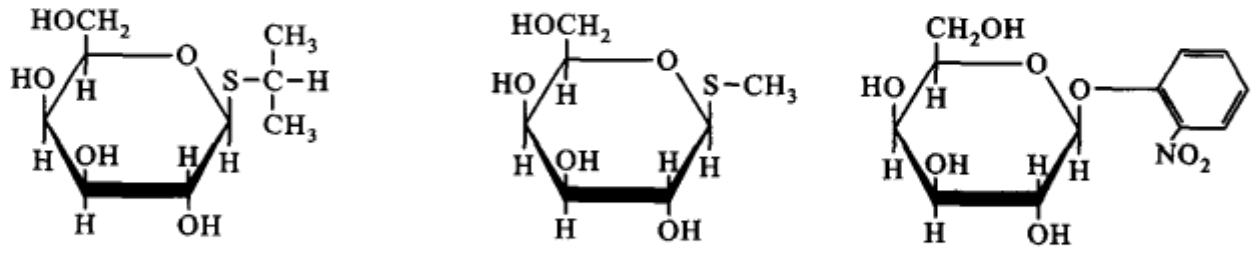



图 7-8 三种乳糖类似物结构式

又称为安慰性诱导物(gratuitous inducer)。

为了证实诱导物的作用是诱导新酶合成,而不是将已存在于细胞中的酶前体转化成有活性的酶,科学家设计了同位素示踪实验。他们把大肠杆菌细胞放在有放射性 $^{35}S$ 标记的氨基酸但没有任何半乳糖苷诱导物的培养基中繁殖几代,然后再将这些带有放射活性的细菌转移到不含 $^{35}S$、无放射性的培养基中,随着培养基中诱导物的加入,β-半乳糖苷酶便开始合成。分离纯化β-半乳糖苷酶,发现这种酶无 $^{35}S$ 标记,说明酶的合成不是由前体转化而来的,而是加入诱导物后新合成的。

## 7.2.2 操纵子模型及其影响因子

1961 年 Jacob 和 Monod 发表 *lac* 操纵子负控诱导模式时,生物界反应之强烈,可与 Watson 和 Crick 在 1953 年发表 DNA 双螺旋模型相媲美。他们还根据基因调控模式预言了 mRNA 的存在,直接导致了后者的发现,还启动了对于氨基酸密码子的实验研究并使整个分子遗传学体系得以迅速建成和完善。

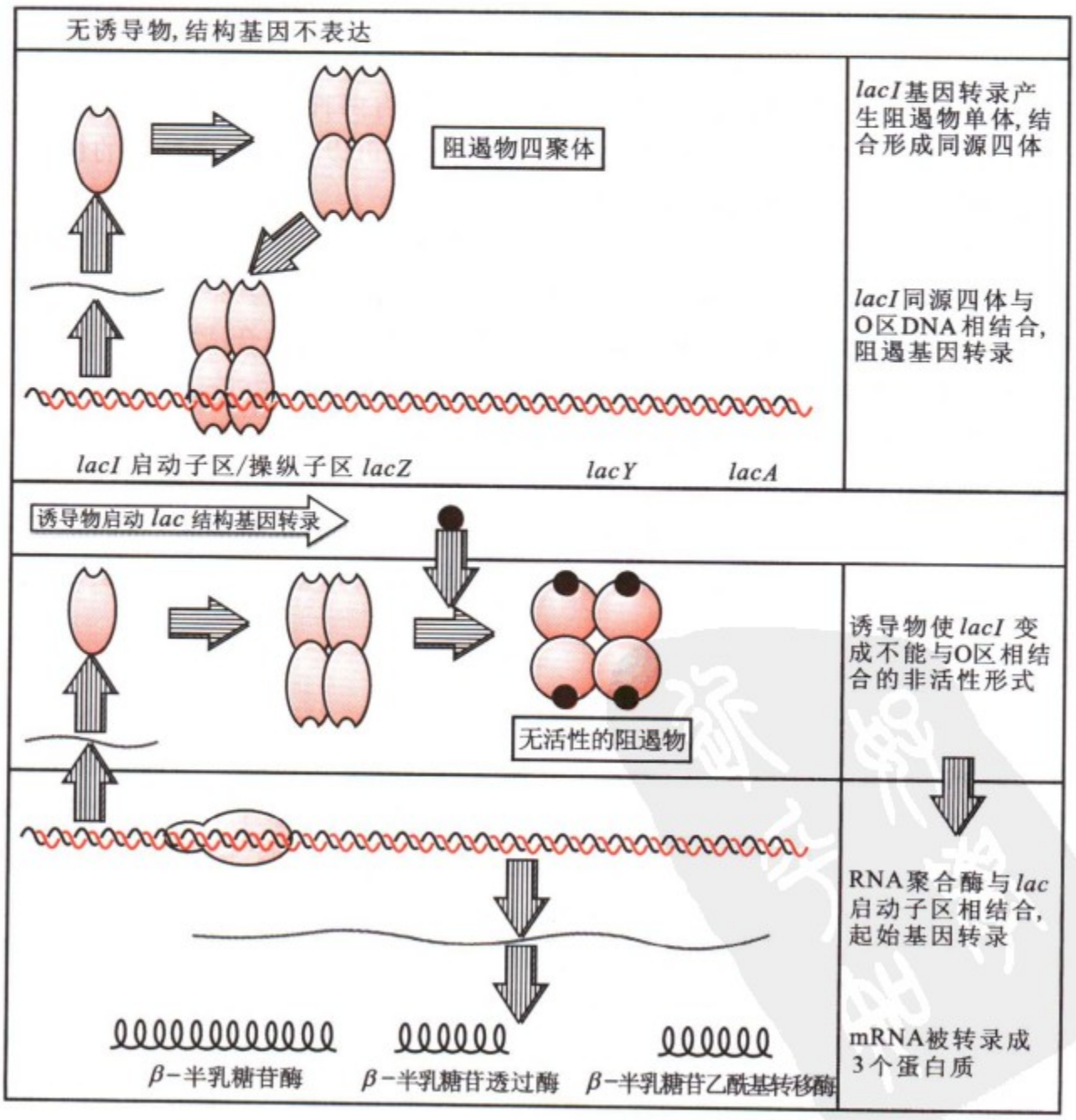



图 7-9 *lac* 操纵子的调控模型

Jacob 和 Monod 认为诱导酶（他们当时称为适应酶）现象是个基因调控问题,而且可以用实验方法进行研究,他们通过大量实验及分析,加上后人的研究成果,使得乳糖操纵子的控制模型被人们广泛接受(图 7-9),其主要内容如下:

① *Z*、*Y*、*A* 基因的产物由同一条多顺反子的 mRNA 分子所编码。

② 该 mRNA 分子的启动区(P)位于阻遏基因(*I*)与操纵区(O)之间,不能单独起始半乳糖苷酶和透过酶基因的高效表达。

③ 操纵区是 DNA 上的一小段序列(仅为 26bp),是阻遏物的结合位点。

④ 当阻遏物与操纵区相结合时,*lac* mRNA 的转录起始受到抑制。

⑤ 诱导物通过与阻遏物结合,改变它的三维构象,使之不能与操纵区相结合,从而激发 *lac* mRNA 的合成。这就是说,有诱导物存在时,操纵区没有被阻遏物占据,所以启动子能够顺利起始 mRNA 的转录。

这一简单模型解释了 *lac* 体系及其他负控诱导系统，我们以后会发现这种解释并不是完善的,因为乳糖操纵子中同样也存在着正调控。

1. *lac* 操纵子的本底水平表达

有两个矛盾是操纵子理论所不能解释的。首先,诱导物需要穿过细胞膜才能与阻遏物结合,而转运诱导物需要透过酶,后者的合成又需要诱导。这样我们必须解释第一个诱导物是如何达到细胞内的。只有两种可能,或者一些诱导物可以在透过酶不存在时进入细胞,或者一些透过酶可以在没有诱导物的情况下合成。研究表明,第二种解释是正确的。

其次,人们发现乳糖(葡糖-1,4-半乳糖)并不与阻遏物相结合,真正的诱导物是乳糖的异构体——异构乳糖(葡糖-1,6-半乳糖),而后者是在 β-半乳糖苷酶的催化下由乳糖形成的,因此,乳糖诱导 β-半乳糖苷酶的合成需要有 β-半乳糖苷酶的预先存在。对这一现象的解释同样是：在非诱导状态下有少量的 *lac* mRNA 合成(大约每个世代中有 1~5 个 mRNA 分子),这种合成被称为本底水平的永久型合成(background level constitutive synthesis)。由于阻遏物的结合并不是绝对紧密的,即使在它与操纵基因紧密结合时,也会偶尔掉下来;这时启动子的障碍被解除,RNA 聚合酶开始转录。这种现象以每个细胞周期 1~2 次的概率发生。

2. 大肠杆菌对乳糖的反应

假设细菌生长在以甘油为碳源的培养基中，按照 *lac* 操纵子本底水平的表达,每个细胞内可有几个分子的 β-半乳糖苷酶和乳糖透过酶。当加入乳糖,在单个透过酶分子的作用下,少量乳糖分子进入细胞,又在单个 β-半乳糖苷酶分子的作用下转变成异构乳糖。某个异构乳糖与结合在操纵区上的阻遏物相结合并使后者失活而离开操纵区,开始了 *lac* mRNA 的生物合成。这些 mRNA 分子编码了大量的 β-半乳糖苷酶和乳糖透过酶,其结果导致乳糖大量涌入细胞。多数乳糖被降解成为葡萄糖和半乳糖,但还有许多乳糖被转变成异构乳糖,然后与细胞内的阻遏物结合(虽然合成速度低,但是阻遏物仍继续合成,因此必须有足够的异构乳糖才能使细胞维持去阻遏状态)，阻遏物的失活促使 mRNA 高速合成,进一步提高了透过酶和 β-半乳糖苷酶的浓度(图 7-10)。降解产生的葡萄

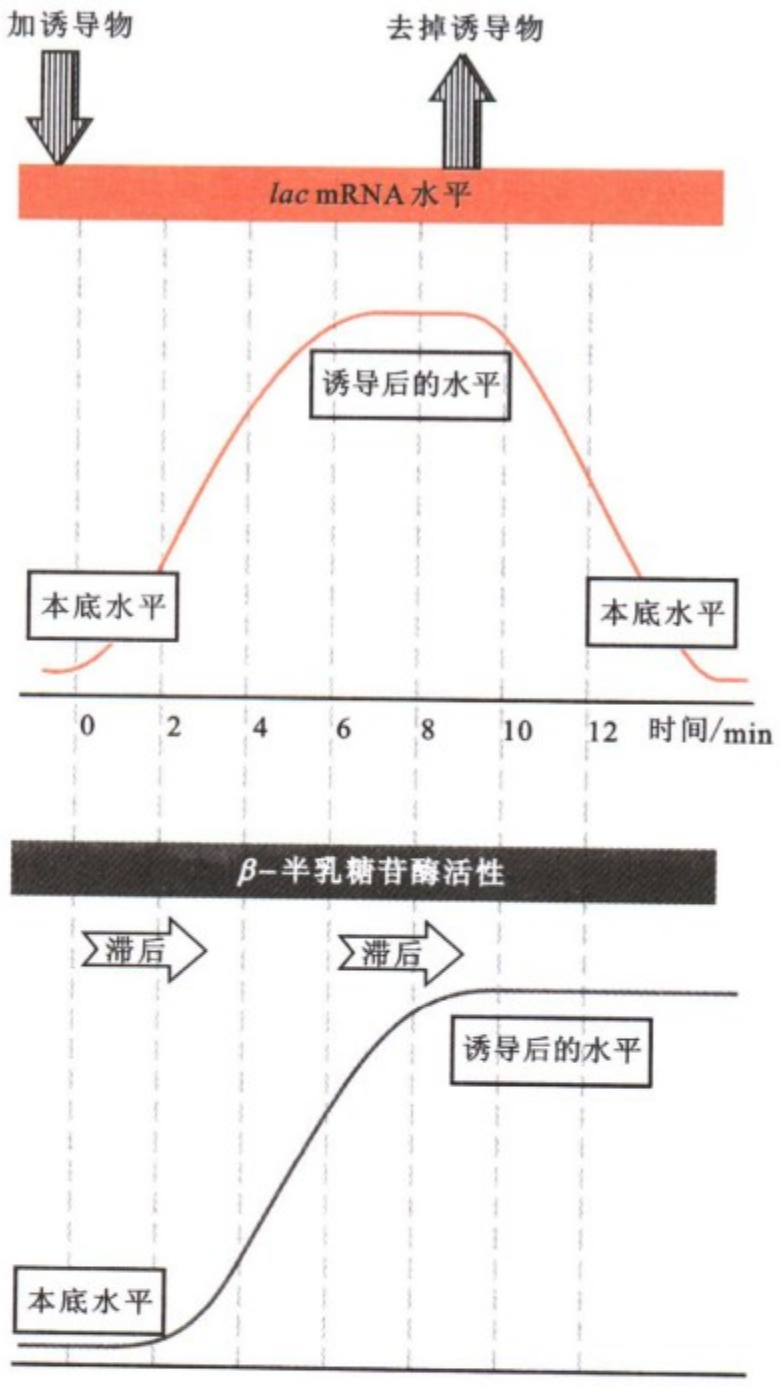



图 7-10 培养基中有无诱导物对 *lac* mRNA 及 β-半乳糖苷酶活性的影响

糖被细胞用作碳源和能源，降解产生的半乳糖则被另一套酶体系转变成葡萄糖，这些酶的合成也是可诱导的。

一旦培养基和细胞中的所有乳糖都被消耗完毕，由于阻遏物仍在不断地被合成，有活性的阻遏物浓度将超过异构乳糖的浓度，使细胞重新建立起阻遏状态，导致 *lac* mRNA 合成被抑制。在细菌中，多数 mRNA 的半衰期只有几分钟，所以在不到一个世代的生长期内，*lac* mRNA 几乎从细胞内消失。β-半乳糖苷酶及透过酶的合成也趋向停止，这些蛋白质虽然很稳定，但其浓度随着细胞的分裂而不断被稀释。值得注意的是，如果在原有的乳糖被撤去之后的一个世代中再加入乳糖，乳糖可以立即开始降解，因为此时细胞内仍有一定浓度的透过酶和β-半乳糖苷酶。

3. 阻遏物 *lac* Ⅰ基因产物及功能

现已查明,*lac* 操纵子阻遏物 mRNA 是在弱启动子控制下永久型合成的,该阻遏蛋白有 4 个相同的亚基,每个亚基均含有 347 个氨基酸残基,并能与 1 分子IPTG 结合。未经分级分离的细胞提取物的结合能力大约为每细胞结合 20~40 个 IPTG 分子,因此推测每个细胞中有 5~10 个阻遏物分子。观察发现在 $I^-$ 突变株细胞提取物中缺少阻遏物,从而证明与 IPTG 结合的物质确实是蛋白质。

当 *I* 基因由弱启动子突变成强启动子,细胞内就不可能产生足够的诱导物来克服阻遏状态,整个 *lac* 操纵子在这些突变体中就不可诱导。

4. 葡萄糖对 *lac* 操纵子的影响

β-半乳糖苷酶在乳糖代谢中的作用是把前者分解成葡萄糖及半乳糖(半乳糖将在 gal 操纵子的作用下再转化成葡萄糖)。如果将葡萄糖和乳糖同时加入培养基中,大肠杆菌在耗尽外源葡萄糖之前不会诱发 *lac* 操纵子(图 7-11)。研究发现,葡萄糖对 *lac* 操纵子表达的抑制是间接的。例如,有一个大肠杆菌突变体,它在糖酵解途径中不能将葡糖-6-磷酸转化为下一步代谢中间物,这些细菌的 *lac* 基因能在葡萄糖存在时被诱导合成。所以,我们说不是葡萄糖而是它的某些降解产物抑制 *lac* mRNA 的合成,科学上把葡萄糖的这种效应称为代谢物阻遏效应(catabolite repression)。

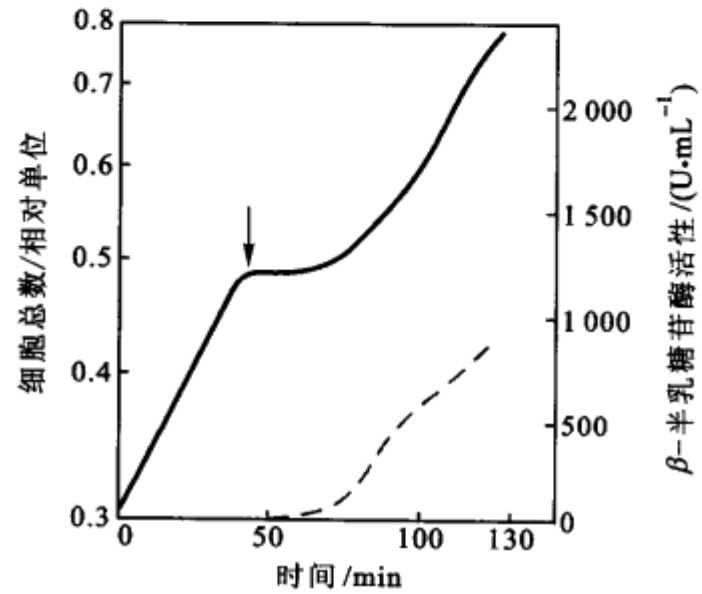



图 7-11 葡萄糖抑制了 *lac* 操纵子基因的表达

当葡萄糖和乳糖同时存在于培养基中时,*lac* 启动子表达受阻,没有 β-半乳糖苷酶活性;当葡萄糖消耗完以后(图中箭头处),细胞内 cAMP 浓度增加,β-半乳糖苷酶活性增加,一度停止生长的细胞又恢复分裂。

5. cAMP 与代谢物激活蛋白

广泛存在于动、植物组织中的 cAMP(图 7-12)是在腺苷酸环化酶的作用下由 ATP 转变而来的,在真核生物的激素调节过程中起着重要作用。在大肠杆菌中,cAMP 的浓度受到葡萄糖代谢的调节。如果将细菌放在缺乏碳源的培养基中培养,细胞内 cAMP 浓度就高;如果在含葡萄糖的培养基中培养,cAMP 的浓度就低;如果培养基中只有甘油或乳糖等不进行糖酵解途径的碳源,cAMP 的浓度也会很高,因此推测糖酵解途径中位于葡糖-6-磷酸与甘油之间的某些代谢产物是腺苷酸环化酶活性的抑制剂。

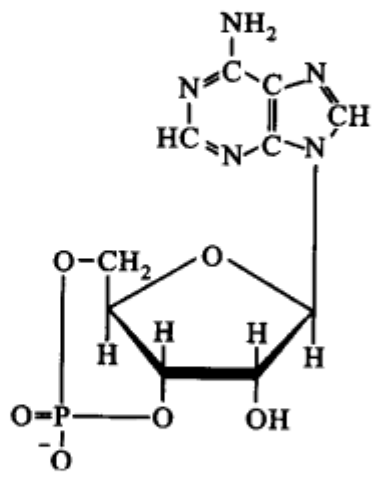



图 7-12 3′,5′-cAMP 的化学结构式

大肠杆菌中的代谢物激活蛋白,由 *Crp* 基因编码,能与 cAMP 形成复合物。凡 *Crp* 及腺苷酸环化酶基因突变的细菌都不能合成 *lac* mRNA,CRP 和 cAMP 都是

合成 *lac* mRNA 所必需的。事实上，cAMP-CRP 复合物是激活 *lac* 的重要组成部分，细菌对于此复合物的需要是独立的，与阻遏体系无关，转录必须有cAMP-CRP 复合物结合在 DNA 的启动子区域上。因为 *Crp* 基因和环化酶基因的突变细菌即使同时也是 $I^-$或 $O^c$ 突变，都不能合成出 *lac* mRNA，所以，我们认为 cAMP-CRP 是一个不同于阻遏物的正调控因子，而 *lac* 操纵子的功能是在这两个相互独立的调控体系作用下实现的。图 7-13 是存在于 *gal*，*lac* 和 *ara* 三个操纵子上游启动子区 CRP-cAMP 结合位点的分析。

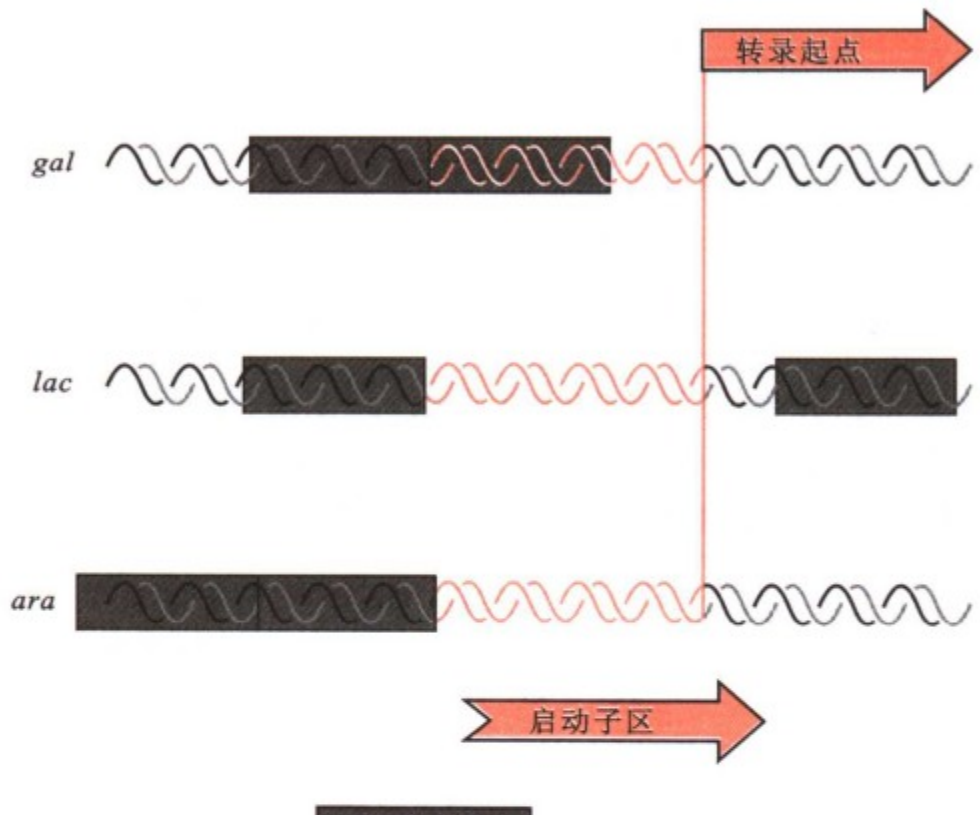



图 7-13 *gal*，*lac* 和 *ara* 操纵子上游启动子区与 CRP-cAMP 结合位点的相对位置分析

此外，半乳糖、麦芽糖、阿拉伯糖、山梨醇等在降解过程中均转化生成葡萄糖或糖酵解途径中的其他中间产物，所以，这些糖代谢中有关的酶都是由可诱导的操纵子控制的。只要有葡萄糖存在，这些操纵子就不表达，被称为降解物敏感型操纵子（catabolite sensitive operon），实验证明这些操纵子都是由 cAMP-

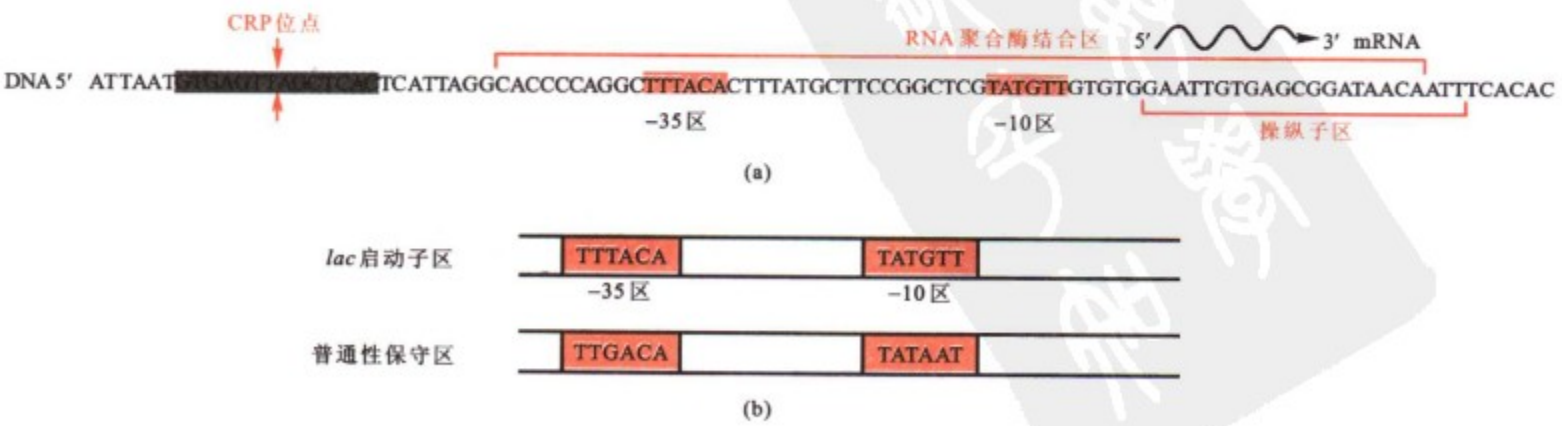
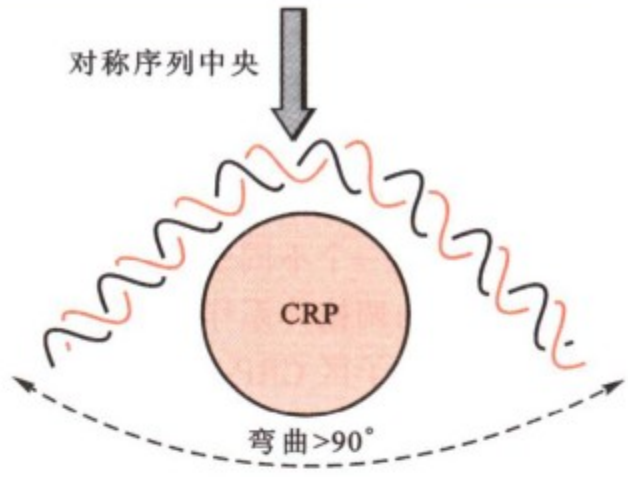



图 7-14 大肠杆菌 *lac* 操纵子启动区 CRP-cAMP 及-35 区，-10 区序列分析

CRP 调节的。对大量操纵子启动区的研究表明,CRP-cAMP 结合位点存在序列特异性(图 7-14)。

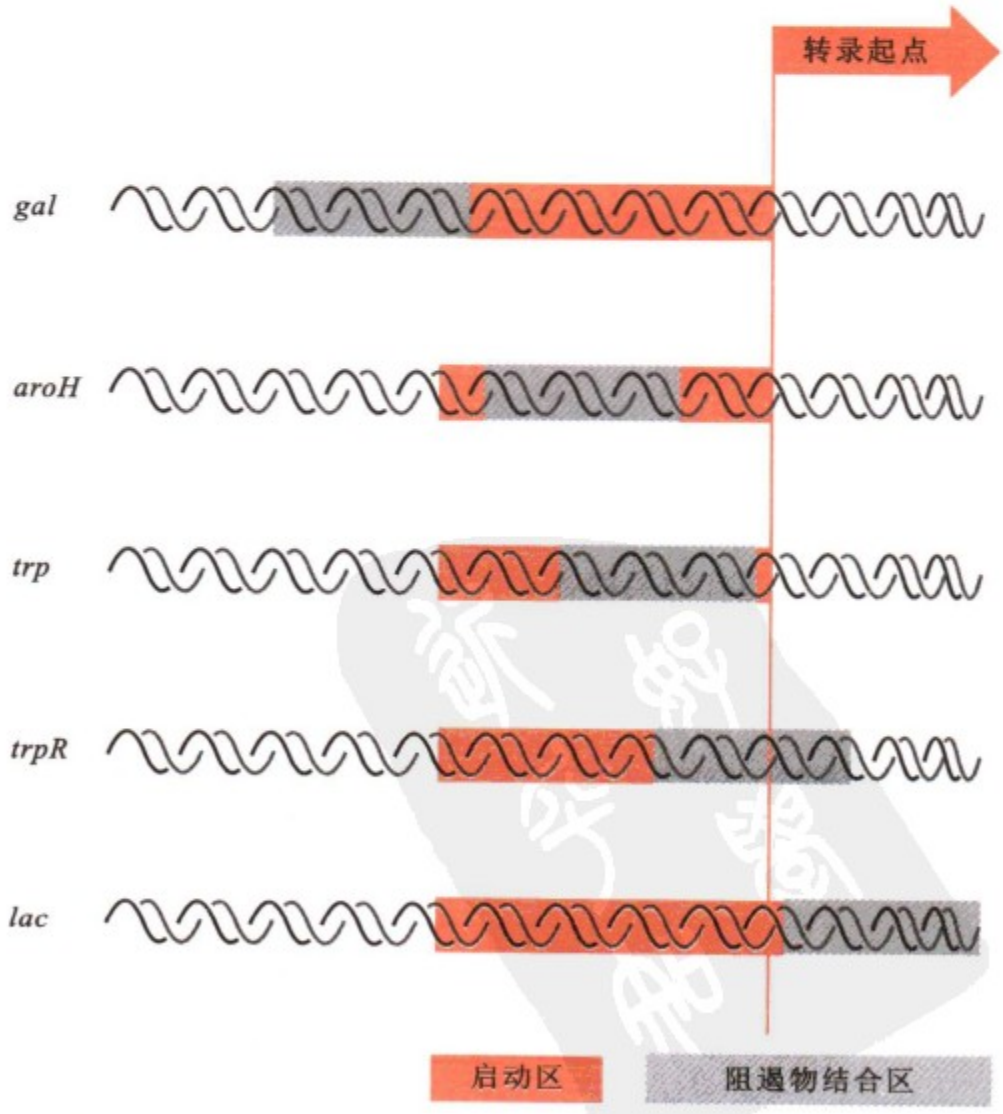
cAMP-CRP 复合物与启动子区的结合是 *lac* mRNA 合成起始所必需的,因为这个复合物结合于启动子上游,能使 DNA 双螺旋发生弯曲,有利于形成稳定的开放型启动子-RNA 聚合酶结构(图 7-15)。而阻遏物则是一个抗解链蛋白(antimelting protein),阻止形成开放结构,从而抑制 RNA 聚合酶的功能。




▌图 7-15 CRP-cAMP 与上游 DNA 位点结合后造成模板 DNA 沿对称序列中央发生了大于 90°的弯曲

### 7.2.3 *lac* 操纵子 DNA 的调控区域——P 区和 O 区

已经分离得到含有 *lac* 操纵子的 DNA 片段,发现 P 区(即启动子区)一般是从 *I* 基因结束到 mRNA 转录起始位点下游 5~10bp,而 O 区(即阻遏物结合区)位于-7~+28 位,该区的碱基序列有对称性,其对称轴在+11 处。图 7-16 分别列




▌图 7-16 不同操纵子启动区及 O 区相对位置比较

举了 *gal*,*aroH*,*trp*,*trpR* 和 *lac* 五个操纵子中 P 区及 O 区的相对位置。实验表明，阻遏物与 O 区的结合影响了 RNA 聚合酶与启动区结合形成转录起始复合物的效率。

### 7.2.4 *lac* 操纵子中的其他问题

1. *A* 基因及其生理功能

*A* 基因存在于 *lac* 操纵子中，编码了 β-半乳糖苷乙酰基转移酶。虽然乳糖本身不能被乙酰化，但这个酶能够把乙酰基从供体上转移到半乳糖苷分子上，这具有什么生理意义呢？

自然界中存在着许多种能被 β-半乳糖苷酶降解的半乳糖苷分子，它们的分解产物往往不能被进一步代谢，很容易在体内积累。这是非常有害的，因为不少半乳糖苷衍生物在高浓度时是细胞正常生长的抑制物。研究发现，当有 IPTG 存在时，$I^-O^cA^-$ 大肠杆菌的生长速度显著慢于相应的 $A^+$ 株系，两组细胞的差异仅仅在于后者能将 IPTG 乙酰化，而乙酰化的 IPTG 不能被 β-半乳糖苷酶所分解。所以，*A* 基因虽然不在乳糖降解中起作用，但它抑制了 β-半乳糖苷酶产物的有害衍生物在细胞内积累，在生物进化中是有意义的。

2. *lac* 基因产物数量上的比较

在一个完全被诱导的细胞中，β-半乳糖苷酶、β-半乳糖苷透过酶及 β-半乳糖苷乙酰基转移酶的拷贝数比例为 1:0.5:0.2。作为数百万年进化演变的结果，这个比例在一定程度上反映了以 β-半乳糖苷作为唯一碳源时细胞的需要。不同的酶在数量上的差异是由于在翻译水平上受到调节所致。这种调节有以下两种方式：

① *lac* mRNA 可能与翻译过程中的核糖体相脱离，从而终止蛋白质链的翻译。这种现象发生的频率取决于每一个后续的 AUG 密码子再度起始翻译的概率。因此，就存在着一个从 mRNA 的 5′端到 3′端的蛋白质合成梯度。对于大多数多顺反子 mRNA 分子来说，这种现象具有普遍性。

② 在 *lac* mRNA 分子内部，*A* 基因比 *Z* 基因更易受内切酶作用而发生降解，因此，在任何时候 *Z* 基因的完整拷贝数要比 *A* 基因多。事实上，对于某个活跃表达的多顺反子操纵子，该 mRNA 所编码的各种蛋白质的相对浓度都由每一个顺反子的翻译频率所决定。

3. 操纵子的融合与基因工程

操纵子在自然条件下可能发生融合，典型的例子是 *lac* 操纵子与负责嘌呤合成的 *pur* 操纵子的偶联。尽管 *pur* 操纵子以完全不同于 *lac* 操纵子的方式进行调节，但它在染色体上位于 *lac* 操纵子沿转录方向的下游，中间只隔了一个控制细胞对 T6 噬菌体敏感性的 *tsx* 基因。科学家从 $tsx^sZ^+$ 细胞中分离到 $tsx^RZ^-$ 突变体，研究该突变体的 DNA 序列，发现它缺失了 *lac* 操纵子的 *Z* 基因、整个 *tsx* 基因以及 *pur* 操纵基因和 *pur* 启动子的一部分。这一突变使 *lac Z* 基因缺失终止序列，当转录作用从 *lac* 启动子处开始后，就合成出由一部分 *Z* 基因编码区和整个 *pur* 基因转录生成的 mRNA 分子。因此，在这个体系中，*pur* 操纵子被“嫁接”

到 *lac* 启动子上,形成融合基因。这一现象给了生物学家以很大的启发,因为 *lac* 启动子是一个很强的启动子,通过它可以使较弱启动子的转录增强,从而增加蛋白质的合成量。到目前为止,基因融合技术已经成为基因工程操作中应用最广泛的技术。

## 7.3 色氨酸操纵子与负控阻遏系统

由于 trp 体系参与生物合成而不是降解,它不受葡萄糖或 cAMP-CRP 的调控。色氨酸的合成主要分 5 步完成,有 7 个基因参与整个合成过程(图 7-17)。*trpE* 和 *trpG* 编码邻氨基苯甲酸合酶,*trpD* 编码邻氨基苯甲酸磷酸核糖转移酶,*trpF* 编码异构酶,*trpC* 编码吲哚甘油磷酸合酶,*trpA* 和 *trpB* 则分别编码色氨酸合酶的 α 和 β 亚基。研究发现,在大肠杆菌等许多细菌中,*trpE* 和 *trpG* 融合成一个功能基因,*trpC* 和 *trpB* 也融合成一个基因,产生具有双重功能的蛋白质。除上述 7 个基因之外, 细菌中还有 *pabA*,*pabB* 基因分别编码 ADC 合酶的两个亚基,*pabC* 基因编码 ADC 裂解酶,*ubiC* 基因编码分枝酸裂解酶,*entC* 基因编码异构分枝酸合酶(图 7-18),这些基因产物都在不同程度上参与了分枝酸代谢,与色氨酸合成有一定关系。

*trpE* 基因是第一个被翻译的基因,和 *trpE* 紧邻的是启动子区和操纵区。另外,前导区和弱化子区分别定名为 *trpL* 和 *trpa*(不是 trpA)。*trp* 操纵子中产生阻遏物的基因是 *trpR*,该基因距 *trp* 基因簇很远。后者位于大肠杆菌染色体图上 25 min 处,而前者则位于 90 min 处。在位于 65 min 处还有一个 *trpS*(色氨酸 tRNA 合成酶),它与携带有色氨酸的 $tRNA^{Trp}$ 共同参与 *trp* 操纵子的调控作用(图 7-19)。*trp* 操纵子的转录调控包括阻遏系统和弱化系统,下面分别予以介绍。

### 7.3.1 *trp* 操纵子的阻遏系统

trpR 基因突变常引起 trp mRNA 的永久型合成, 该基因产物因此被称为辅阻遏蛋白(aporepressor)。除非培养基中有色氨酸,否则这个辅阻遏蛋白不会与操纵区结合。辅阻遏蛋白与色氨酸相结合形成有活性的阻遏物,与操纵区(图 7-20)结合并关闭 *trp* mRNA 转录。

研究发现,这个系统中的效应物分子是色氨酸,是由 *trp* 操纵子所编码的生物合成途径的末端终产物。当培养基中色氨酸含量较高时,它与游离的辅阻遏蛋白相结合,并使之与操纵区 DNA 紧密结合;当培养基中色氨酸供应不足时,辅阻遏物失去色氨酸并从操纵区上解离,*trp* 操纵子去阻遏。

### 7.3.2 弱化子与前导肽

1. 弱化子

随着对色氨酸操纵子的深入研究,发现有些现象与以阻遏作为唯一调节机制的观点不一致,例如,在色氨酸高浓度和低浓度下观察到 *trp* 操纵子的表达水

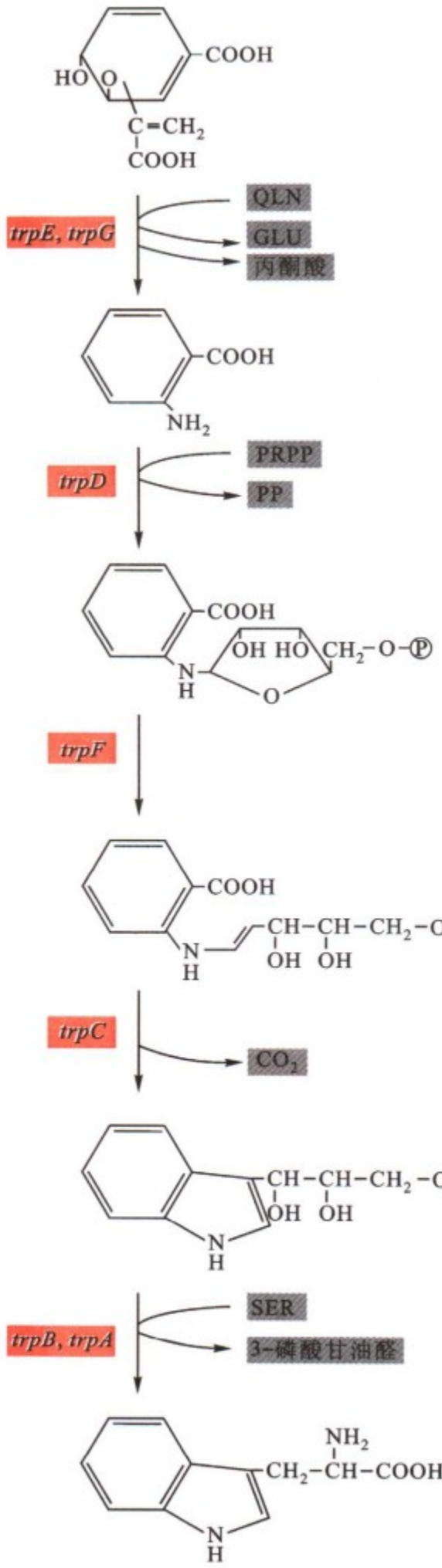



图 7-17 细菌中色氨酸生物合成途径

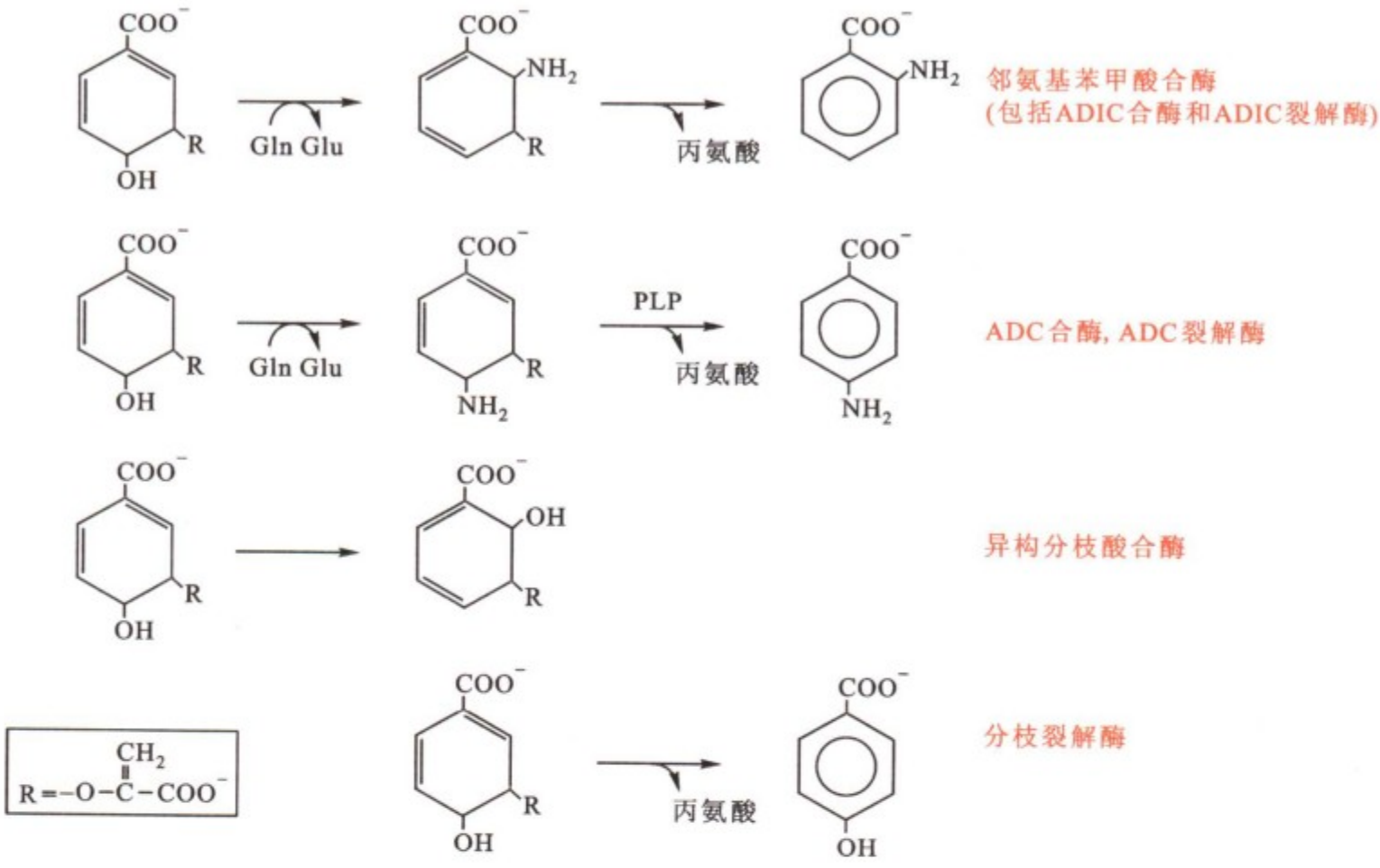



图 7-18 细菌中参与分枝酸代谢的主要酶系统

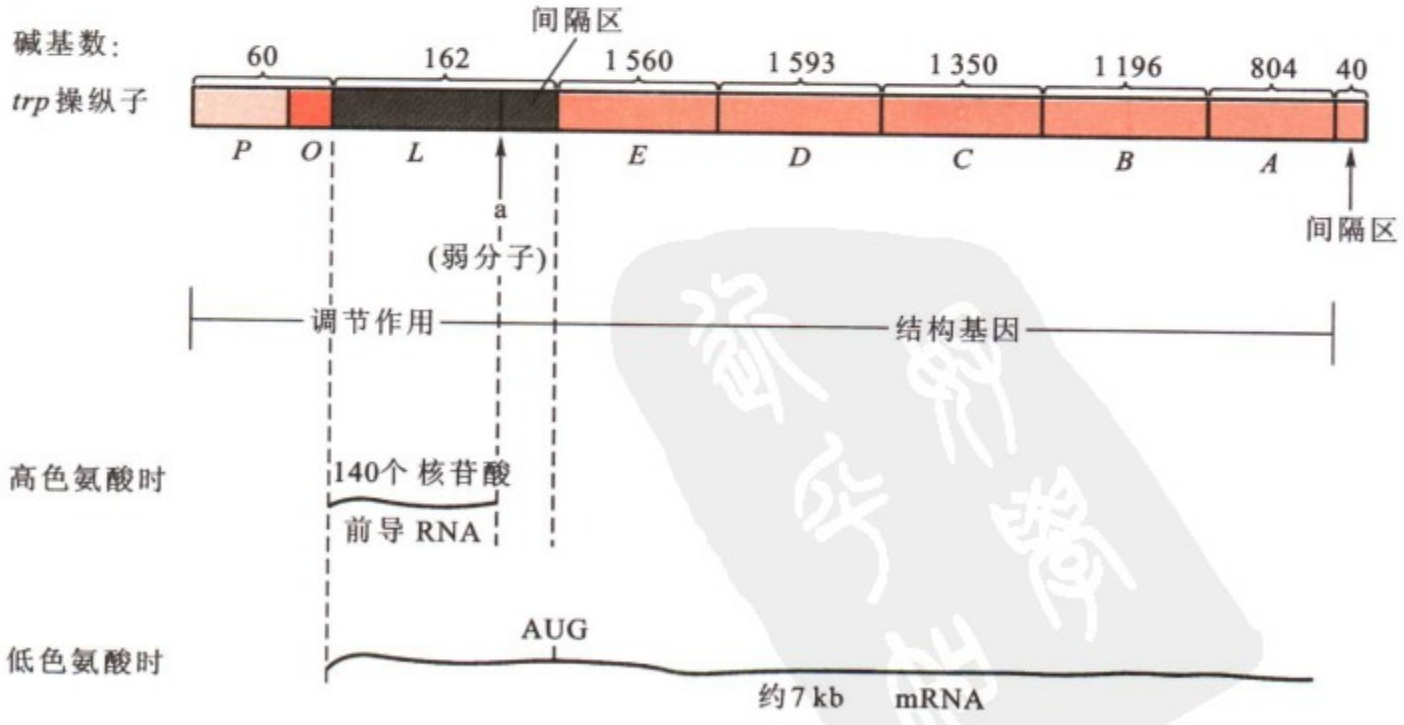



图 7-19 大肠杆菌中的 *trp* 操纵子

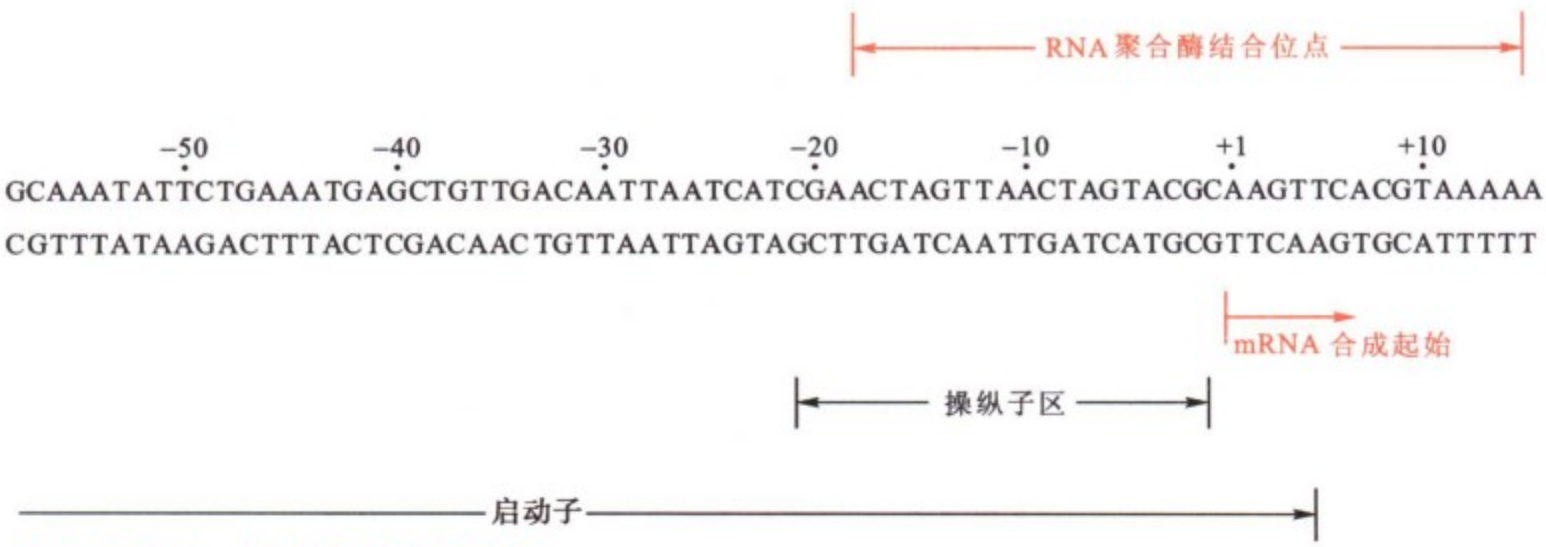



图 7-20 *trp* 操纵区 DNA 序列特征

平相差约 600 倍,然而阻遏作用仅使转录降低 70 倍。此外,使阻遏物失活突变不能完全消除色氨酸对 *trp* 操纵子表达的影响。显然,*trp* 操纵子的这种调控与阻遏物的控制无关,必然还有其他调控机制。通过缺失突变株的研究发现,这种调控机制就是弱化作用(attenuation)。事实上,阻遏-操纵机制对色氨酸来说只是一个一级开关,控制转录的启动,相当于 *trp* 操纵子的粗调开关。*trp* 操纵子中对应于色氨酸生物合成的还有另一个系统进行细调控并决定已经启动的转录是否继续下去。在这个系统中,酶的浓度根据氨基酸浓度的变化而受到调节。这个细微调控是通过转录达到第一个结构基因之前的过早终止来实现的,细胞中色氨酸的浓度是实现过早终止的根本原因。

在 *trp* mRNA 5′端 *trpE* 基因的起始密码前有一个长 162 bp 的 mRNA 片段被称为前导区,其中 123~150 位碱基序列如果缺失,*trp* 基因表达可提高 6~10 倍,而且无论是在阻遏细胞内还是在永久性突变的细胞内都是这样。研究发现,当 mRNA 合成起始以后,除非培养基中完全没有色氨酸,转录总是在这个区域终止,产生一个仅有 140 个核苷酸的 RNA 分子,终止 *trp* 基因转录,这就是 123~150 区序列缺失会提高 *trp* 基因表达的原因。因为转录终止发生在这一区域,并且这种终止是被调节的,这个区域就被称为弱化子。研究引起终止的 mRNA 碱基序列(图 7-21),发现该区 mRNA 通过自我配对可以形成茎-环结构,有典型的终止子特点。

2. 前导肽

各种实验表明衰减作用需要负载 $tRNA^{Trp}$ 参与,这意味着前导序列的某些部分被翻译了。分析前导序列发现,它包括起始密码子 AUG 和终止密码子 UGA;如果翻译起始于 AUG,应该产生一个含有 14 个氨基酸的多肽。这个假设的多肽(还未实际观察到)被称为前导肽。有证据表明事实上可能存在这个多肽,因为在该 AUG 位点 5′上游端有一个核糖体结合位点,而前导肽编码区与 *trpE* 基因 5′端融合可产生一种融合蛋白质,该蛋白质具有与前导肽同样的 N 端序列。

前导序列具有一个非常有意义的特点,在其第 10 和第 11 位上有相邻的两

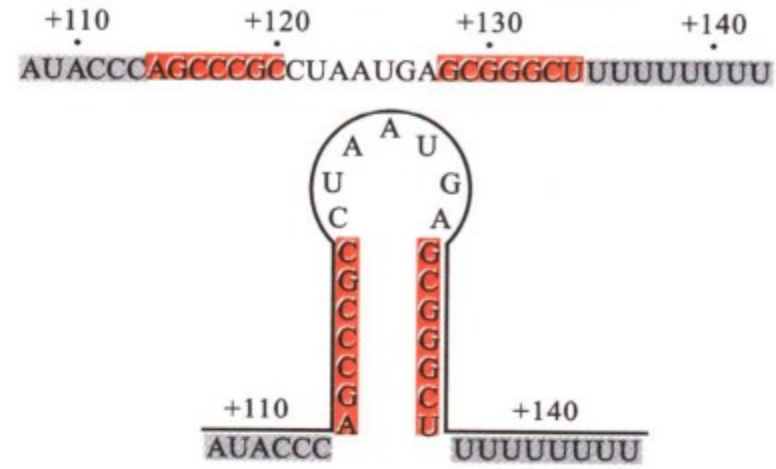



图 7-21 *trp* 弱化子 mRNA 的终止区

个色氨酸密码子。这一点很重要,因为组氨酸操纵子中,也具有弱化子,也具有一个类似的能编码前导肽的碱基序列，此序列中含有 7 个相邻的组氨酸密码子。苯丙氨酸操纵子中同样存在弱化子结构,其前导序列中也有 7 个苯丙氨酸密码子。这些密码子参与了 *trp* 及其他操纵子中的转录弱化机制。

3. mRNA 前导区的序列分析

*trp* 前导区的碱基序列已经全部测定完成，引人注目的是其中 4 个分别以 1、2、3 和 4 表示的片段能以两种不同的方式进行碱基配对 (图 7-22)，有时以 1-2 和 3-4 配对,有时只以 2-3 方式互补配对。RNaseT1 降解实验(此酶不能水解配对的 RNA)表明,纯化的 trp 前导序列中确有 1-2 和 3-4 的配对方式,由此

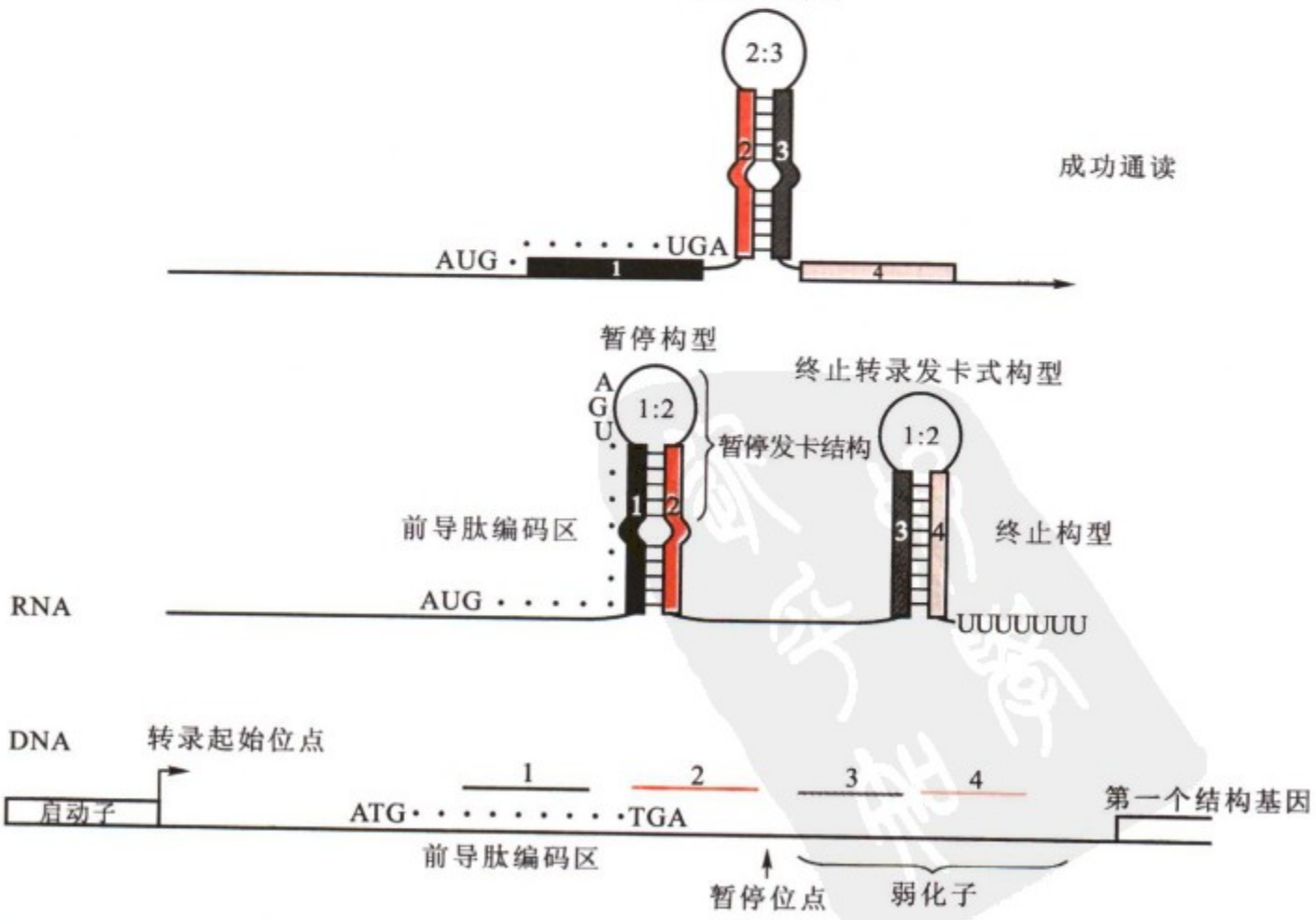



图 7-22 原核生物 RNA 中前导序列变构引起转录的终止或弱化

定位的 3–4 配对区正好位于终止密码子的识别区，当这个区域发生破坏自我配对的碱基突变时有利于转录的继续进行(图 7–23)。

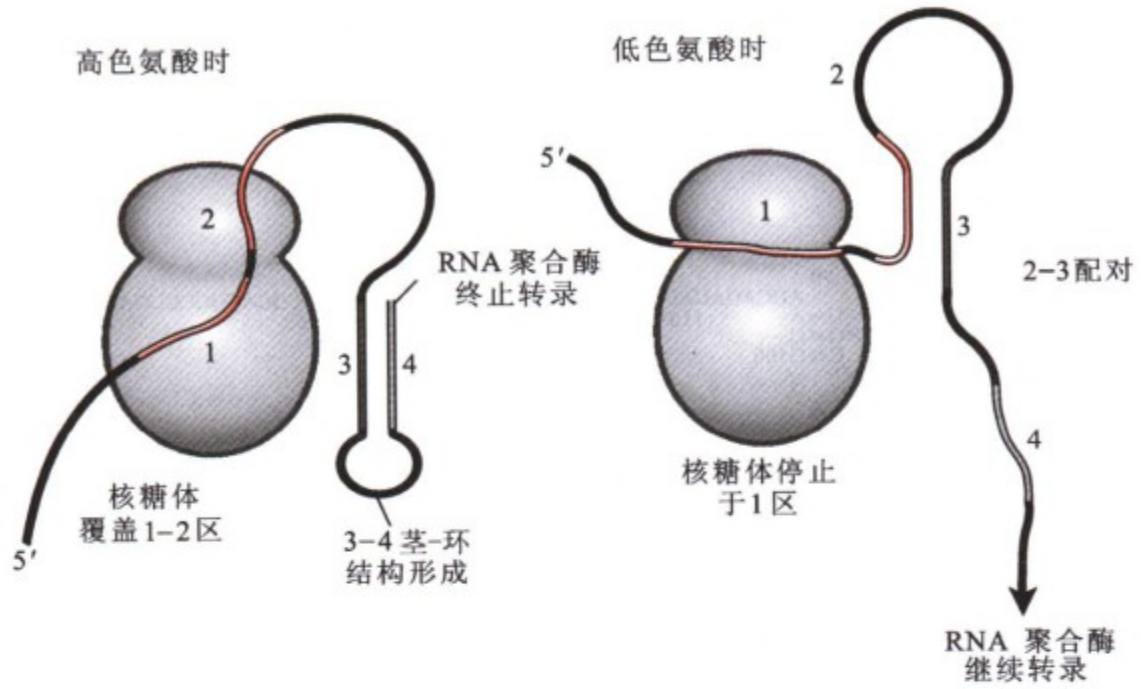



图 7–23 *trp* 操纵子的转录与翻译调控

4. 转录的弱化作用

转录的弱化理论认为 mRNA 转录的终止是通过前导肽基因的翻译来调节的。因为在前导肽基因中有两个相邻的色氨酸密码子，所以这个前导肽的翻译必定对 $tRNA^{Trp}$ 的浓度敏感。当培养基中色氨酸的浓度很低时，负载有色氨酸的 $tRNA^{Trp}$ 也就少，这样翻译通过两个相邻色氨酸密码子的速度就会很慢，当 4 区被转录完成时，核糖体才进行到 1 区(或停留在两个相邻的 *trp* 密码子处)，这时的前导区结构是 2–3 配对，不形成 3–4 配对的终止结构，所以转录可继续进行，直到将 *trp* 操纵子中的结构基因全部转录。而当培养基中色氨酸浓度高时，核糖体可顺利通过两个相邻的色氨酸密码子，在 4 区被转录之前，核糖体就到达 2 区，这样使 2–3 不能配对，3–4 区可以自由配对形成茎–环状终止子结构，转录停止，trp 操纵子中的结构基因被关闭而不再合成色氨酸(图 7–24)。所以，弱化子对 RNA 聚合酶的影响依赖于前导肽翻译中核糖体所处的位置。已从大肠杆菌和沙门菌中陆续发现了不少具有弱化作用的操纵子，它们都具有前导肽，并且前导肽中都富含该操纵子所合成的那种氨基酸。

### 7.3.3 *trp* 操纵子弱化机制的实验依据

在含有较高浓度色氨酸的培养基中，色氨酰–tRNA 合成酶缺陷型大肠杆菌中 *trp* 操纵子结构基因的表达却并没有完全受到抑制。进一步研究温度敏感突变株 *trpS5* 证实，它所编码的 trp–tRNA 合成酶只在 30℃时有活性，能使色氨酸活化生成 $tRNA^{Trp}$，在 42℃时该酶没有活性。比较野生型和突变型在 42℃和30℃条件下有色氨酸供应时邻氨基苯甲酸合酶(trpE)的活性发现，30℃时，色氨酰–tRNA 合成酶有活性，可生成 $tRNA^{Trp}$，*trp* 操纵子结构基因的表达受抑制；在42℃

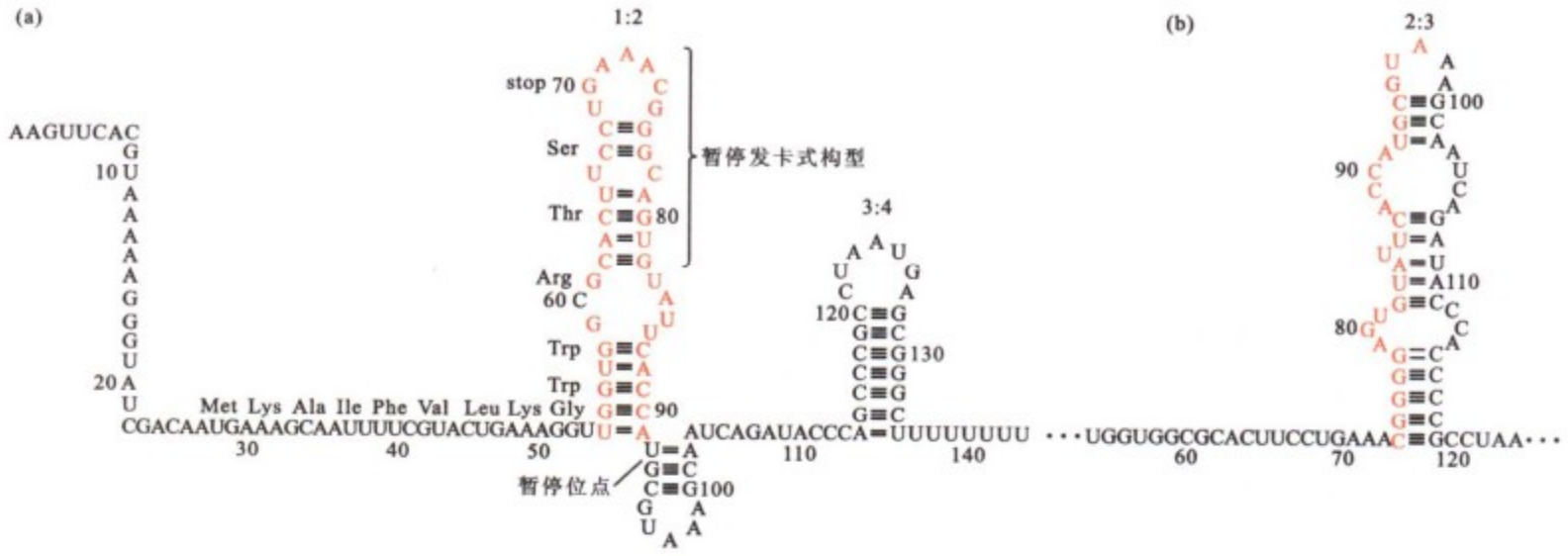



**图 7-24 *trp* 操纵子前导序列的变构与转录的弱化** (a)转录终止构象;(b)抗终止构象。

时,色氨酰-tRNA 合成酶失去活性,不能生成 $tRNA^{Trp}$,*trp* 结构基因表达受抑制的程度就不明显。野生型大肠杆菌色氨酰-tRNA 合成酶不受温度控制,所以 *trp* 基因表达基本不受温度影响。

研究另一个前导区及 *D* 基因缺失的 *trp* 操纵子突变体(trpΔLD102)还发现,该细菌在补充了过量色氨酸的培养基中有高于野生型亲本株系 8~10 倍的色氨酸合成酶活性,这种去阻遏作用不论在 $trpR^+$或 $trpR^-$菌株中都存在(表 7-3)。

**表 7-3 *trpL* 缺失突变株与非缺失菌株色氨酸合成酶活性比较(相对单位)**

| 株系 | $trpR^+$ | $trpR^-$ |
|---|---|---|
| TrpΔED53($trpa^+$) | 0.17 | 11.8 |
| TrpΔLD102($trpa^-$) | 1.82 | 100 |

从表 7-3 中可以看出前导区的调节作用。trpΔED53 的 L 区中不缺失弱化子,而 trpΔLD102L 区中缺失弱化子。缺失前导区后的表达比有前导区的表达要高得多,充分说明 *trp* 操纵子的表达调控除阻遏作用外,还受到前导区的影响,失去了这个因素就失去了一个调控机制。

*trpL29* 是增强终止的突变型,它的前导序列 29 位核苷酸由原来的 G 变为 A,……AUG……→……AUA……,AUG 是前导肽的起始密码,变为 AUA 后便不能起始蛋白质合成,有利于 *trp* mRNA 形成 1-2、3-4 稳定的二级结构,从而增强终止。这一突变型的邻氨基苯甲酸合成酶活性很低,并且对色氨酸不敏感。因为前导肽根本不能翻译下去,即使在色氨酸饥饿条件下仍然不能作出解除终止的反应。

另一个增强终止的突变型是 *trpL75*,其 75 位上的 G 变成了 A。野生型 75 位上的 G 在 2 区内,它可以与 3 区内 118 位上的 C 形成氢键,当 G 变为 A 后,减弱了 2 区和 3 区形成茎-环结构的能力,导致终止增强,*trpL75* 对色氨酸饥饿不能作出反应。

此外,缺失了 3′端 4 个 U 的突变株 ΔtrpLC1419,也导致终止解除,说明 3′端 U 对转录终止起作用。细菌通过弱化作用弥补阻遏作用的不足,因为阻遏作

用只能使转录不起始,而对于已经起始的转录,只能通过弱化作用使之中途停顿下来。阻遏作用的信号是细胞内色氨酸的多少,弱化作用的信号则是细胞内载有色氨酸的 tRNA 的多少。它通过前导肽的翻译来控制转录的进行,在细菌细胞内这两种作用相辅相成,体现着生物体内周密的调控作用。

细菌中为什么需要有弱化子系统呢?一般认为,阻遏物从有活性向无活性的转变速度较慢,需要有一个能更快地做出反应的系统,以保持培养基中适当的色氨酸水平。弱化子能较快地通过抗终止的方法来增加 *trp* 基因表达,迅速提高内源色氨酸浓度。

那么,为什么还要有阻遏体系呢?没有阻遏体系,只有弱化,似乎也可以调节色氨酸酶系的合成。目前认为阻遏物的作用是当有大量外源色氨酸存在时,阻止非必需的先导 mRNA 的合成,它使这个合成系统更加经济。事实上,自然界中存在着不同类型的合成体系。组氨酸操纵子拥有在功能上与 *trp* 操纵子完全相同的弱化子结构,但没有阻遏物,它的表达完全受弱化子调节。

### 7.3.4 trp 操纵子的其他调控机制

大肠杆菌中共有 5 个基因参与色氨酸生物合成,构成色氨酸操纵子。其中 *trpG–D* 和 *trpC–F* 为融合基因,翻译出的多肽具有双重功能。有两个启动子(一个位于操纵子 5′,一个位于 *trpG–D* 编码区,称为内源启动子)控制整个操纵子的转录,能够感受色氨酸和无负载的 $tRNA^{Trp}$ 的浓度变化。

枯草埃希菌(*Bacillus subtilis*)色氨酸操纵子包括 6 个基因,处于一个 12 个基因的大操纵子内。其他基因控制部分芳香族氨基酸的生物合成。*B.subtilis* 还有第 7 个色氨酸合成基因 *trpG*(又称 *pabA*),位于叶酸操纵子中。翻译出来的 TrpG–PabA 多肽链能发挥两个酶的作用,一个用于色氨酸途径,一个用于叶酸途径。*B.subtilis* 的色氨酸操纵子也有两个启动子区。RNA 聚合酶在任何一个启动子区起始转录都会受到色氨酸和无负载 $tRNA^{Trp}$ 的调控(图 7–25)。

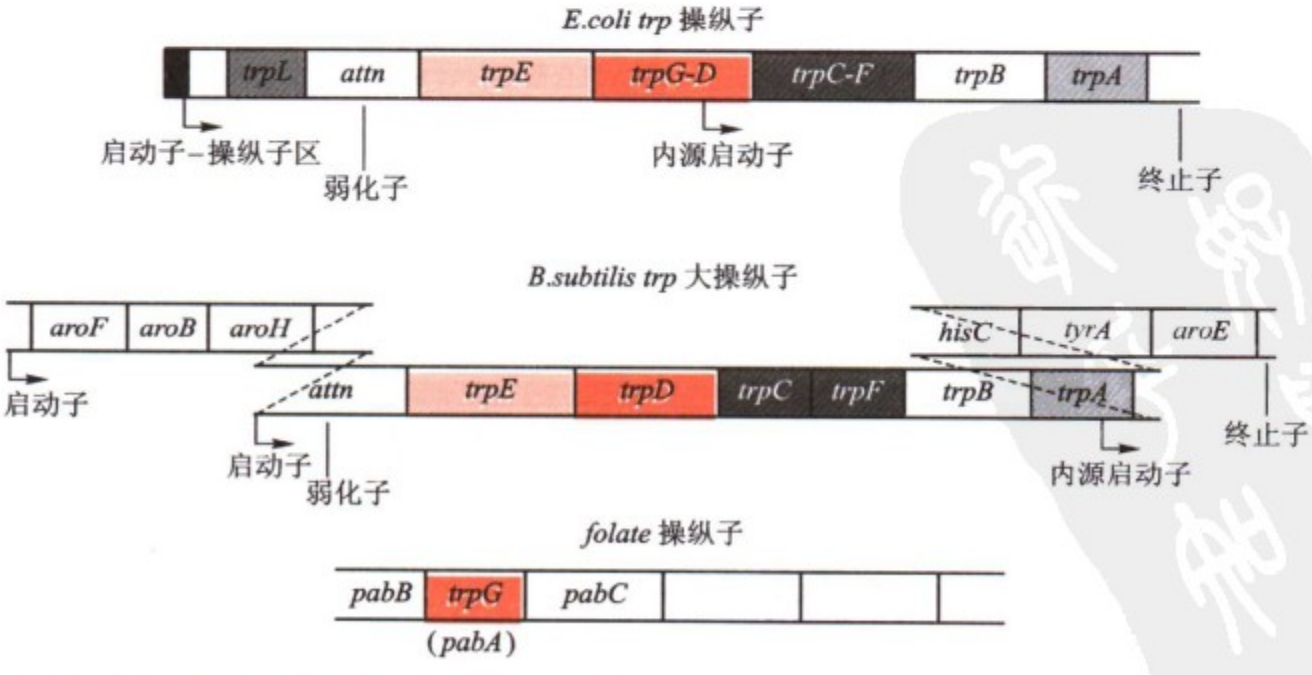



图 7–25 *B.subtilis* 和 *E.coli* 色氨酸操纵子的结构比较

大肠杆菌色氨酸操纵子受到由色氨酸激活的负阻遏蛋白的调节作用。一旦转录越过前导区,开始结构基因的转录,又会受弱化子的调控,感受无负载的 $tRNA^{Trp}$ 的变化,使转录机器在前导区附近停止或者继续结构基因的转录。在 *Bacillus subtilis* 色氨酸操纵子中,色氨酸激活 TRAP 调节蛋白并使之与前导 RNA 相结合,形成 RNA 终止子结构终止转录。而无负载 $tRNA^{Trp}$ 的聚集使 TRAP 蛋白失活,从而激活转录和翻译。另外,当无负载 $tRNA^{Trp}$ 的浓度升高时,由 *rtpA* 基因编码的 AT 蛋白与 TRAP 蛋白结合,抑制该蛋白形成转录终止复合物从而激活色氨酸操纵子的转录(图 7-26)。

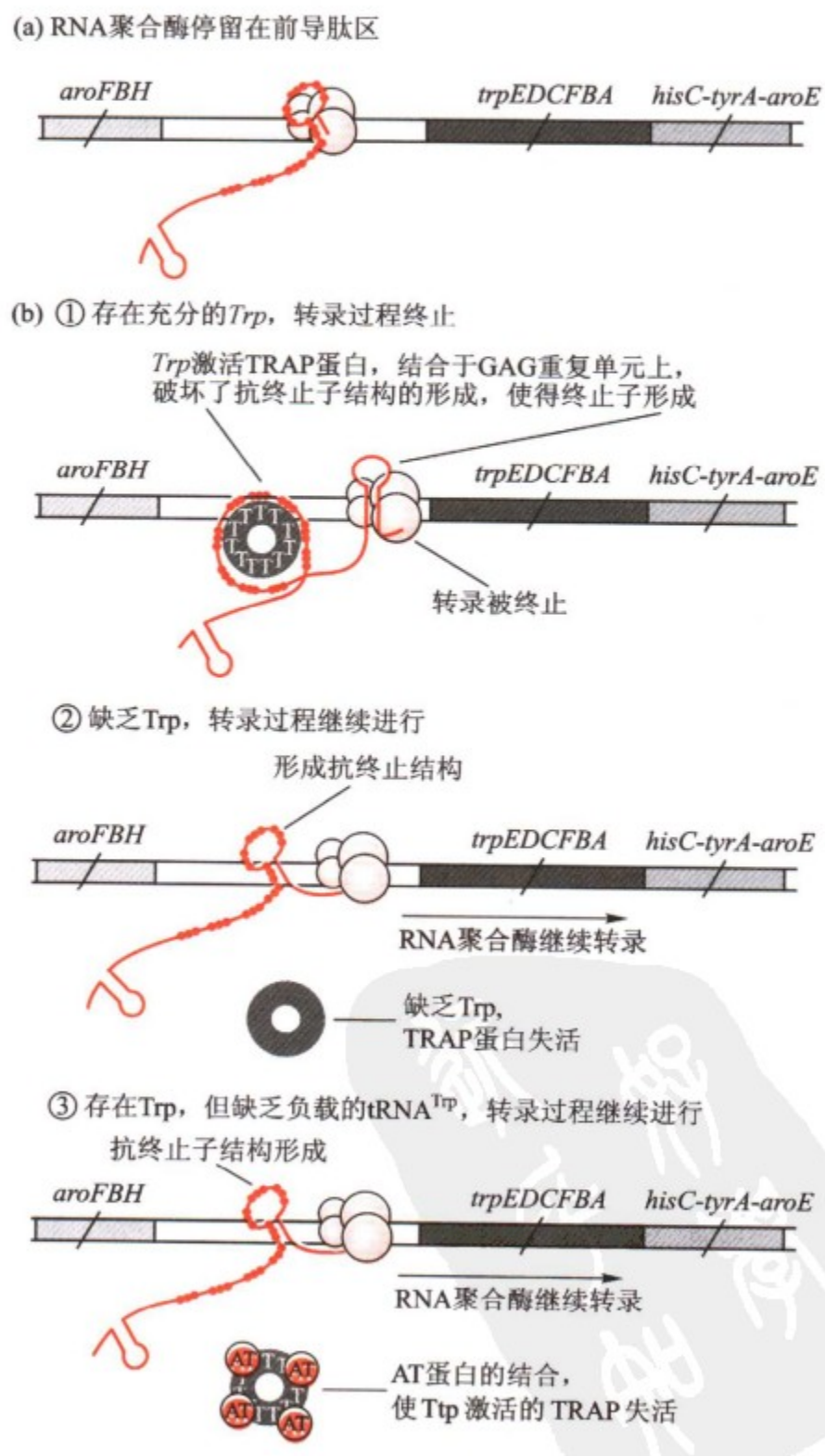



图 7-26 *B.subtilis* 中色氨酸操纵子的调控机制

# 7.4 其他操纵子

## 7.4.1 半乳糖操纵子

大肠杆菌半乳糖操纵子（galactose operon）在大肠杆菌遗传图上位于17 min处，包括3个结构基因：异构酶（UDP-galactose-4-epimerase，galE），半乳糖-磷酸尿嘧啶核苷转移酶（galactose transferase，galT），半乳糖激酶（galactose kinase，galK）。这3个酶的作用是使半乳糖变成葡糖-1-磷酸。半乳糖操纵子的调节基因是*galR*，位于遗传图上55 min处。*gal*操纵子与*lac*操纵子所不同的是*galR*与*galE*、*T*、*K*及操纵区*O*等的距离都很远，而*galR*产物对

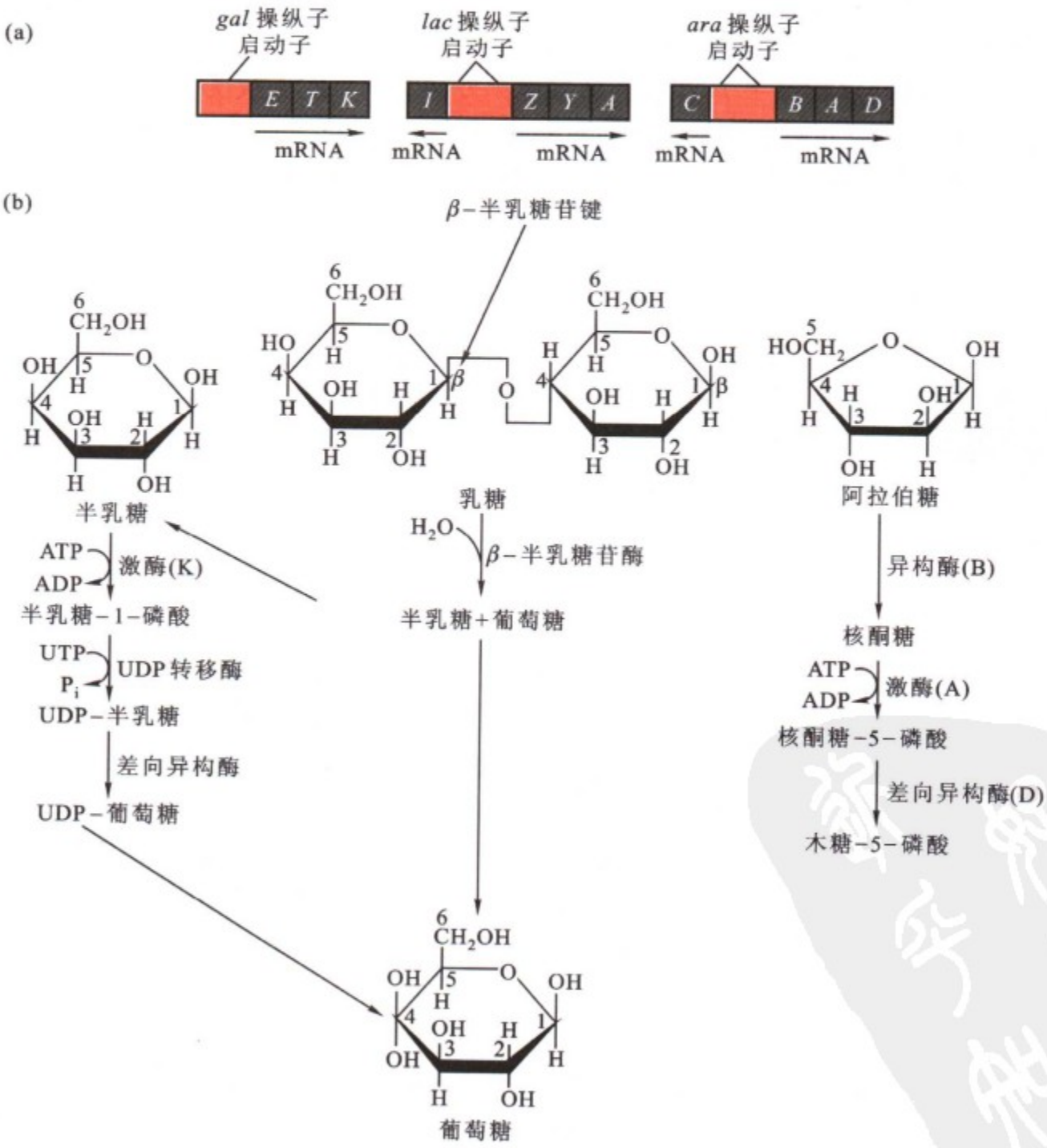



**图 7-27 *gal*操纵子、*lac*操纵子和*ara*操纵子的结构及其代谢途径** (a)操纵子的结构；(b)代谢途径。

*galO* 的作用与 *lacI–lacO* 的作用相同。在 *galR*⁻和 *galO*ᶜ 突变体中，*E*、*T*、*K* 基因得到永久性表达。*gal* 操纵子的诱导物主要是半乳糖。图 7–27 是几个相关联操纵子的代谢途径及基因图谱。半乳糖的跨细胞膜运转需要 *galP* 基因产物的参与。

此外，*gal* 操纵子还有两个特点：①它有两个启动子，其 mRNA 可从两个不同的起始点开始转录；②它有两个 *O* 区，一个在 *P* 区上游–73~–67，另一个在结构基因 *galE* 内部。

1. cAMP–CRP 对 *gal* 启动子的作用

因为半乳糖的利用效率比葡萄糖低，人们猜想葡萄糖存在时半乳糖操纵子不被诱导，但实际上有葡萄糖存在时，*gal* 操纵子仍可被诱导。现已分离到一些突变株，其中一类突变株能在不含葡萄糖的培养基中高水平合成半乳糖代谢酶类(*gal* 结构基因高效表达)；另一类突变株中，*gal* 基因的表达完全依赖于葡萄糖，培养基中如无葡萄糖存在，这些细菌的 *gal* 基因不表达，不合成半乳糖代谢酶类。分析 *gal* 操纵子 *P–O* 区的碱基序列发现有两个相距仅为 5bp 的启动子，*gal* 操纵子可以从两个启动子分别起始基因转录，每个启动子拥有各自的 RNA 聚合酶结合位点 $S_1$ 和 $S_2$(图 7–28)。cAMP–CRP 对从 $S_1$ 和 $S_2$ 起始的转录有不同的作用。

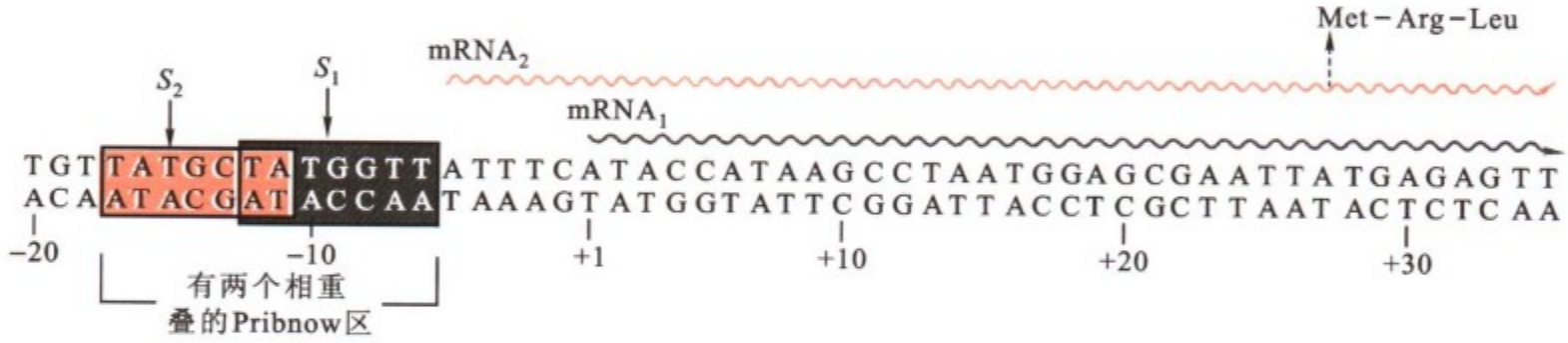



图 7–28 *gal* 操纵子启动区 $S_1$ 和 $S_2$ 序列分析

从 $S_1$ 起始的转录只有在培养基中无葡萄糖时才能顺利进行，RNA 聚合酶与 $S_1$ 的结合需要半乳糖、CRP 和较高浓度的 cAMP。当腺苷环化酶突变(cya⁻)或cAMP 受体蛋白突变(crp⁻)时，*gal* 操纵子不能从 $S_1$ 起始转录。此时如在体外系统中加入 cAMP–CRP，就能够诱发从 $S_1$ 起始的转录。体外转录的实验证明，cAMP–CRP 能够刺激 *gal* 操纵子转录，用含有 *gal* 操纵子调节区的 DNA 片段作模板进行体外转录时，发现 cAMP–CRP 抑制从 $S_2$ 的转录而刺激从 $S_1$ 的转录。当有cAMP–CRP 时，转录从 $S_1$ 开始，当无 cAMP–CRP 时，转录从 $S_2$ 开始。

从 $S_2$ 起始的转录则完全依赖于葡萄糖，高水平的 cAMP–CRP 能抑制由这个启动子起始的转录，因此，cAMP–CRP 在 *gal* 操纵子中所起的调节作用比在 lac 操纵子中更为复杂。大肠杆菌的 *cya*⁻或 *crp*⁻突变型不能利用乳糖，但可以利用半乳糖作为唯一碳源。

有一个*gal* 启动子突变株，不能利用培养基中的半乳糖，若将此突变株再

行突变为 *cya*⁻或 *crp*⁻，细胞就恢复了利用半乳糖的能力。因为第一次突变失去了从 $S_1$ 起始转录的能力，却不影响从 $S_2$ 起始转录的能力。细胞中 cAMP–CRP 的存在抑制了从 $S_2$ 开始的基因转录，使 *gal* 操纵子既不能从 $S_1$ 起始又无法从 $S_2$ 起始转录。双重突变后，无论是 *cya*⁻突变，还是 *crp*⁻突变，都能移去 cAMP–CRP 对 $S_2$ 的阻遏作用，导致 *gal* 基因表达，恢复利用半乳糖作为能源的能力。

一般认为，cAMP–CRP 有利于 RNA 聚合酶–$S_1$ 区复合物形成开链构象，从而起始基因转录。同时，由于 $S_1$ 和 $S_2$ 区的核苷酸部分重叠，这一复合物的存在干扰了 RNA 聚合酶–$S_2$ 复合物的形成，抑制 $S_2$ 起始的基因转录。

2. 双启动子的生理功能

为什么 *gal* 操纵子需要两个转录起始位点？这与半乳糖在细胞代谢中的双重功能有关。半乳糖不仅可以作为唯一碳源供细胞生长，而且与之相关的物质——尿苷二磷酸半乳糖(UDPgal)是大肠杆菌细胞壁合成的前体。生长过程中细胞必须随时合成差向异构酶，以保证尿苷二磷酸的供应。在没有外源半乳糖的情况下，细胞通过半乳糖差向异构酶(galactose epimerase，*galE* 基因产物)的作用由 UDP–葡萄糖合成 UDPgal。因为合成细胞壁过程中对异构酶的需要量很小，本底水平的永久型合成就能够满足生理需要。实际上，*gal* mRNA 的永久型合成水平已高于 *lac* 操纵子所合成 *lac* mRNA 的水平，显然这部分 mRNA 是在没有半乳糖的情况下合成的。

现在设想只有 $S_1$ 一个启动子，那么由于这个启动子的活性依赖于 cAMP–CRP，当培养基中有葡萄糖存在时就不能合成异构酶。假如唯一的启动子是 $S_2$，那么，即使在有葡萄糖存在的情况下，半乳糖也将使操纵子处于充分诱导状态，这无疑是一种浪费。所以，无论从必要性或经济性考虑，都需要一个不依赖于 cAMP–CRP 的启动子($S_2$)进行本底水平的永久型合成，以及一个依赖于 cAMP–CRP 的启动子($S_1$)对高水平合成进行调节。这就是说，只有在 $S_2$ 活性完全被 cAMP–CRP 抑制时，调控作用才是有效的。

### 7.4.2 阿拉伯糖操纵子

阿拉伯糖(arabinose)是另一个可以为代谢提供碳源的五碳糖。在大肠杆菌中阿拉伯糖的降解需要 3 个基因：*araB*、*araA* 和 *araD*，它们形成一个基因簇，简写为 araBAD。它们因分别编码 3 个酶：*araB* 基因编码核酮糖激酶(ribulokinase)，araA 编码 L–阿拉伯糖异构酶 (L–arabinose isomerase)，araD 编码 L–核酮糖–5–磷酸–4–差向异构酶(L–ribulose–5–phosphate–4–epimerase)。在 *gal* 操纵子中，基因的排列序列就是代谢途径中酶的作用序列，而这个操纵子却不同，阿拉伯糖的代谢是以 *araA*、*araB*、*araD* 的序列进行的。另外还有两个负责将阿拉伯糖运入细胞的基因，即 *araE* 和 *araF*，它们位于远离这个基因簇的地方，*araE* 和 *araF* 分别编码一个膜蛋白和一个位于细胞壁与细胞膜之间的阿拉伯糖结合蛋白。

与 *araBAD* 相邻的是一个复合的启动子区域、两个操纵区($O_1$,$O_2$)和一个调

节基因 *araC*，这个调节基因的性质与已介绍过的 *lac* 及 *gal* 操纵子中的调节基因的性质全然不同。AraC 蛋白同时显示正、负调节因子的功能。在 *lac* 和 *gal* 操纵子中，阻遏蛋白只能作为负调节因子，而 cAMP-CRP 蛋白只能是正调节因子。*araBAD* 和 *araC* 基因的转录是分别在两条链上以相反的方向进行的。在标准的遗传学图谱上，*araBAD* 基因簇从启动子 $P_{BAD}$ 开始向左进行转录，而 *araC* 基因则是从 $P_c$ 向右转录(图 7-29)。

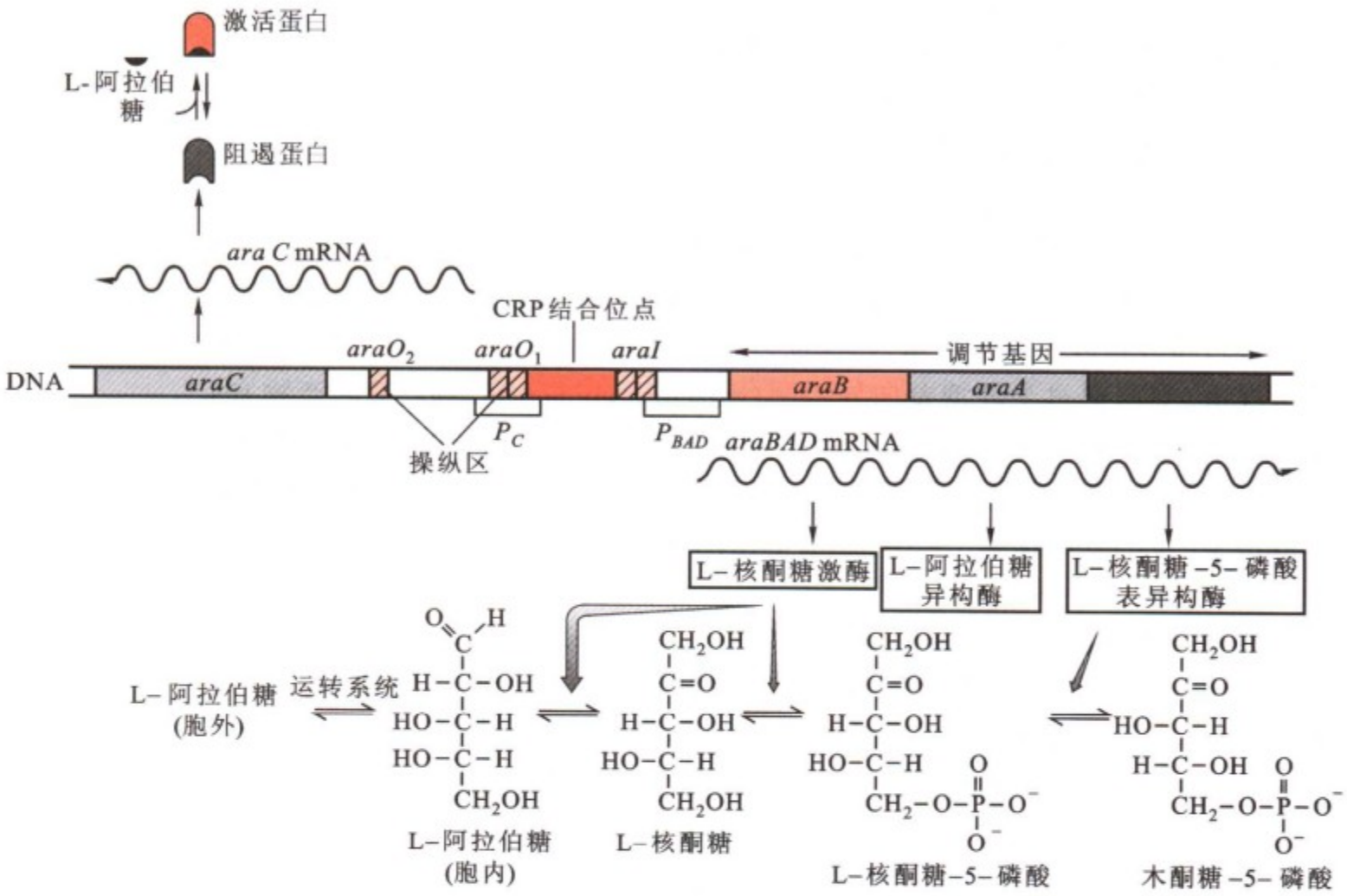



图 7-29 *ara* 操纵子上游调控区序列与功能分析

*ara* 操纵子也是可诱导的，阿拉伯糖本身就是诱导物。在野生型操纵子中，只有阿拉伯糖存在时才转录出 *araBAD* mRNA，而有葡萄糖存在时则不转录。腺苷酸环化酶缺陷型和 *crp*⁻突变株也不形成 *araBAD* mRNA，说明从 $P_{BAD}$ 起始的转录过程也需要 cAMP-CRP。

1. araC 蛋白的正、负调节作用

AraC 蛋白既是 *ara* 操纵子的正调节蛋白，又是其负调节蛋白(图 7-29)。我们来分析以下实验结果。首先，当 *araC* 基因内发生点突变或缺失突变，会产生不能合成 *araBAD* mRNA 的 araC⁻突变株。其次，构建基因型为 F′*araC*⁺/*araC*⁻的局部二倍体细胞后发现这类细胞中 *ara* 操纵子可以被诱导，说明 *araC*⁻等位基因是隐性的。如果 *araC*⁻突变体的产物是突变型的阻遏物，那么，*araC*⁻对于 *araC*⁺来说至少应有部分显性。

遗传学上的其他证据如 *araC* 基因的琥珀突变，也说明 *araC* 基因产物是一

种蛋白质，这一点又通过 AraC 蛋白的纯化得到了进一步证实。所以说，*araBAD* mRNA 的合成需要一个有功能的 AraC 蛋白，这个蛋白与诱导物阿拉伯糖结合后，才诱导 *araBAD* 基因表达(图 7–30)。

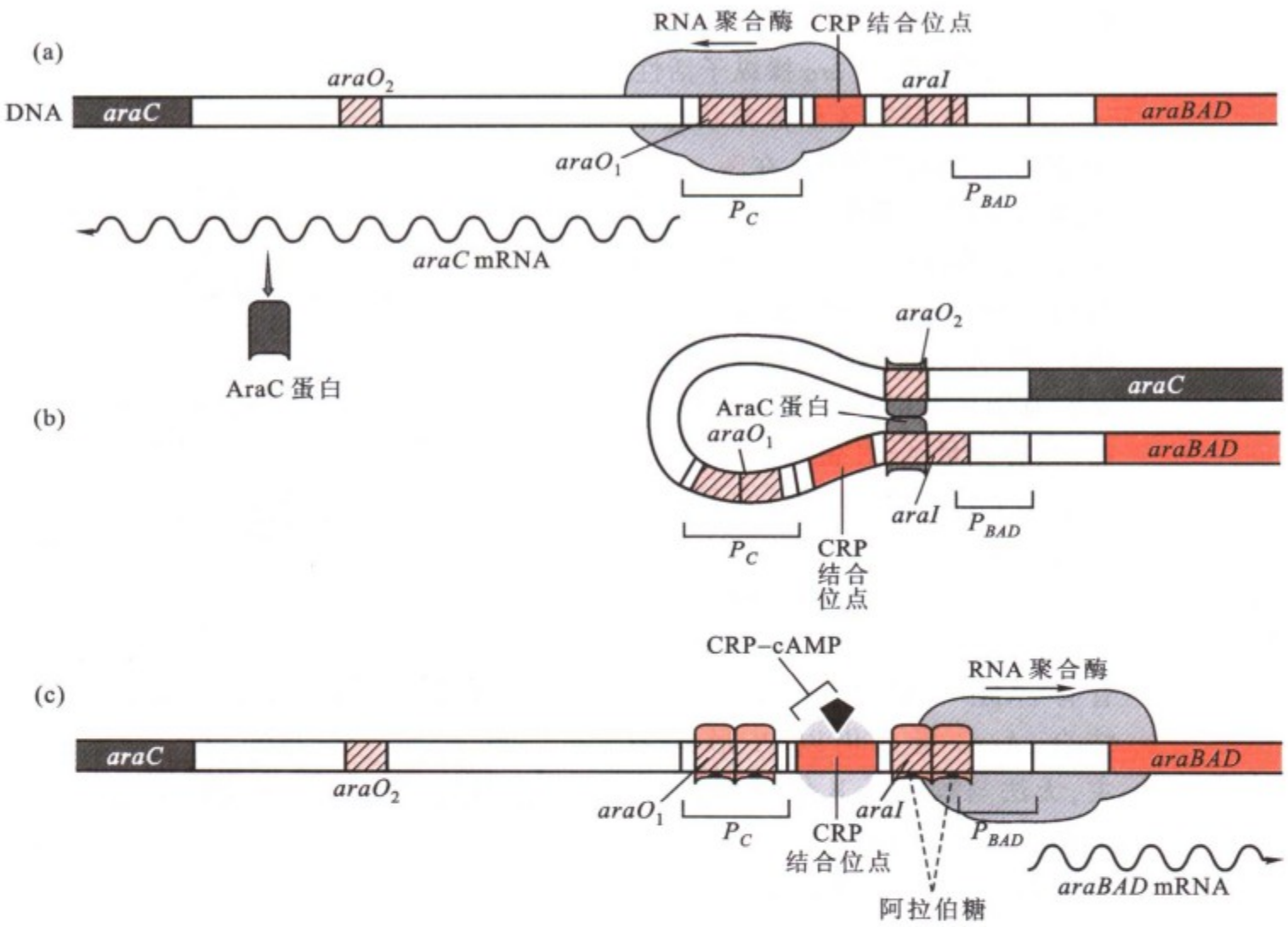



**图 7–30 *ara* 操纵子的表达调控** (a) 当细胞内没有 AraC 蛋白时，由 $P_c$ 启动子起始 *araC* 基因转录；(b) 当体系中葡萄糖水平较高，阿拉伯糖水平较低时，AraC 蛋白与操纵区 $O_2$ 以及 *araI* 诱导因子结合区上半区相结合，形成 DNA 回转结构，*ara BAD* 基因不表达；(c) 当体系中有阿拉伯糖但无葡萄糖存在时，AraC 蛋白与阿拉伯糖相结合，改变构象成为激活蛋白，AraC 蛋白同源二聚体分别与 $araO_1$ 和 *araI* 区相结合，DNA 回转结构被破坏，RNA 聚合酶在 AraC 蛋白和 CRP–cAMP 的共同作用下起始 $P_{BAD}$ 所调控的结构基因表达。

有实验证明，当 AraC 蛋白以正调控因子作用时，起始转录还需要 cAMP–CRP 的共同参与。在第一个实验里，人们发现当反应体系中只含有 *ara* 操纵子 DNA、RNA 聚合酶、AraC 蛋白及 cAMP–CRP 中的一个成分，*araBAD* mRNA 的转录不能起始，说明只有 AraC 蛋白和 cAMP–CRP 协同作用才能起始 *araBAD* 的转录。在第二个实验里，人们用电镜直接观察 *ara* 操纵子 DNA 分子(这种技术可以看到结合了 RNA 聚合酶的 DNA 分子)发现，除非 AraC 蛋白和 cAMP–CRP 同时存在，RNA 聚合酶一般不能与该操纵子 DNA 结合。

2. AraC 蛋白的两种形式

AraC 蛋白作为 $P_{BAD}$ 活性正、负调节因子的双重功能是通过该蛋白的两种异

构体来实现的。一般认为,Pr 是起阻遏作用的形式,可以与现在尚未鉴定的类操纵区位点相结合,而 Pi 是起诱导作用的形式,它通过与 $P_{BAD}$ 启动子结合进行调节。推测 Pr 和 Pi 这两种结构处于一种平衡之中。在没有阿拉伯糖时,Pr 形式占优势;一旦有阿拉伯糖存在,它就能够与 AraC 蛋白结合,使平衡趋向于 Pi 形式。

3. 营养状况对 *ara* 操纵子活性的影响

图 7-31 是在各种营养状态下 *ara* 操纵子的表达情况。在图 7-31a 中,培养基含有葡萄糖,cAMP-CRP 没有与操纵区位点相结合,AraC 蛋白处于 Pr 形式并与操纵区 A 位点结合,RNA 聚合酶不能与 $P_c$ 结合,*araC* 基因虽然仍有转录,但受到抑制,只有少量 AraC 蛋白形成,整个系统几乎处于静止状态。它虽然不是完全关闭,但是尽可能地关闭了。图 7-31b 中没有葡萄糖,也没有阿拉伯糖,因为没有诱导物,尽管有 cAMP-CRP 与操纵区位点相结合,AraC 蛋白仍以 Pr 形式为主,无法与操纵区 B 位点结合,无 *araBAD* mRNA 转录。图 7-31c 中无葡萄糖,有阿拉伯糖,大量 *araC* 基因产物以 Pi 形式存在,并分别与操纵区 B、A 位点相结合,在 cAMP-CRP 的共同作用下,*araC* 和 *araBAD* 基因大量表达,操纵子充分激活。

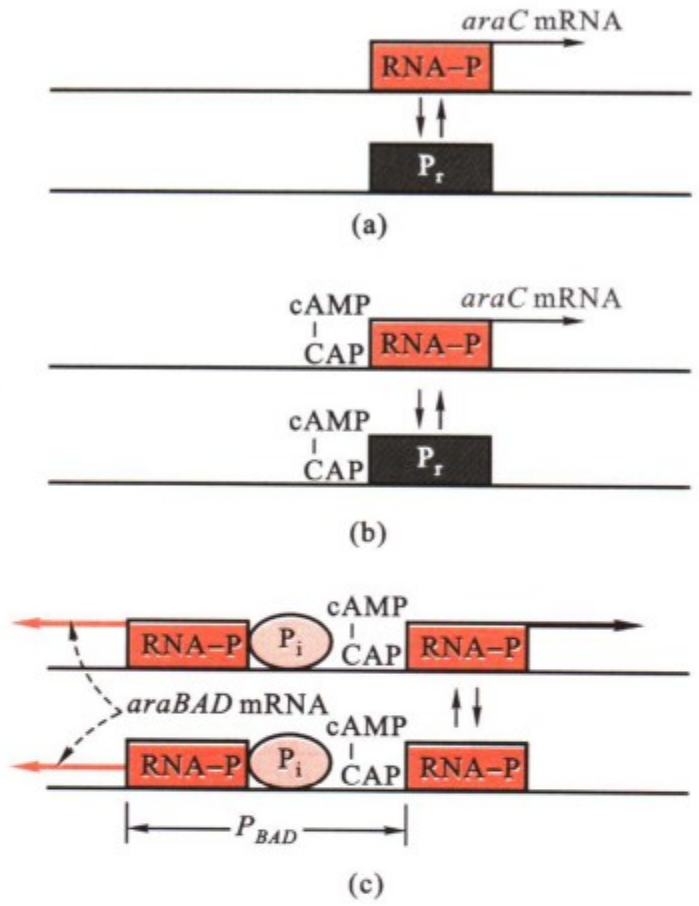



**图 7-31 各种营养状态下 *ara* 操纵子的表达调控(RNA-P 代表 RNA 聚合酶)** (a)有葡萄糖,无阿拉伯糖;(b)无葡萄糖,无阿拉伯糖;(c)无葡萄糖,有阿拉伯糖。

## 7.4.3 阻遏蛋白 LexA 的降解与细菌中的 SOS 应答

当细菌 DNA 遭到破坏时(如受到紫外线照射),细菌细胞内会启动一个被称为 SOS 的诱导型 DNA 修复系统。研究发现,参与 SOS DNA 修复系统的许多基因虽然分散在染色体的各个部位,但都同时受 LexA 阻遏蛋白的抑制,平时表达水平很低。SOS 体系的诱导表达过程其实就是把 LexA 阻遏蛋白从这些基因的上游调控区移开的过程。

一般情况下,*recA* 基因表达并不完全受 lexA 阻遏,所以,每个细胞中可有 1 000 个 RecA 蛋白单体分散在细胞质中。当 DNA 严重受损时,DNA 复制被中断,单链 DNA 缺口数量增加,RecA 与这些缺口处单链 DNA 相结合,被激活成为蛋白酶,将 LexA 蛋白切割成没有阻遏和操纵区 DNA 结合活性的两个片段,

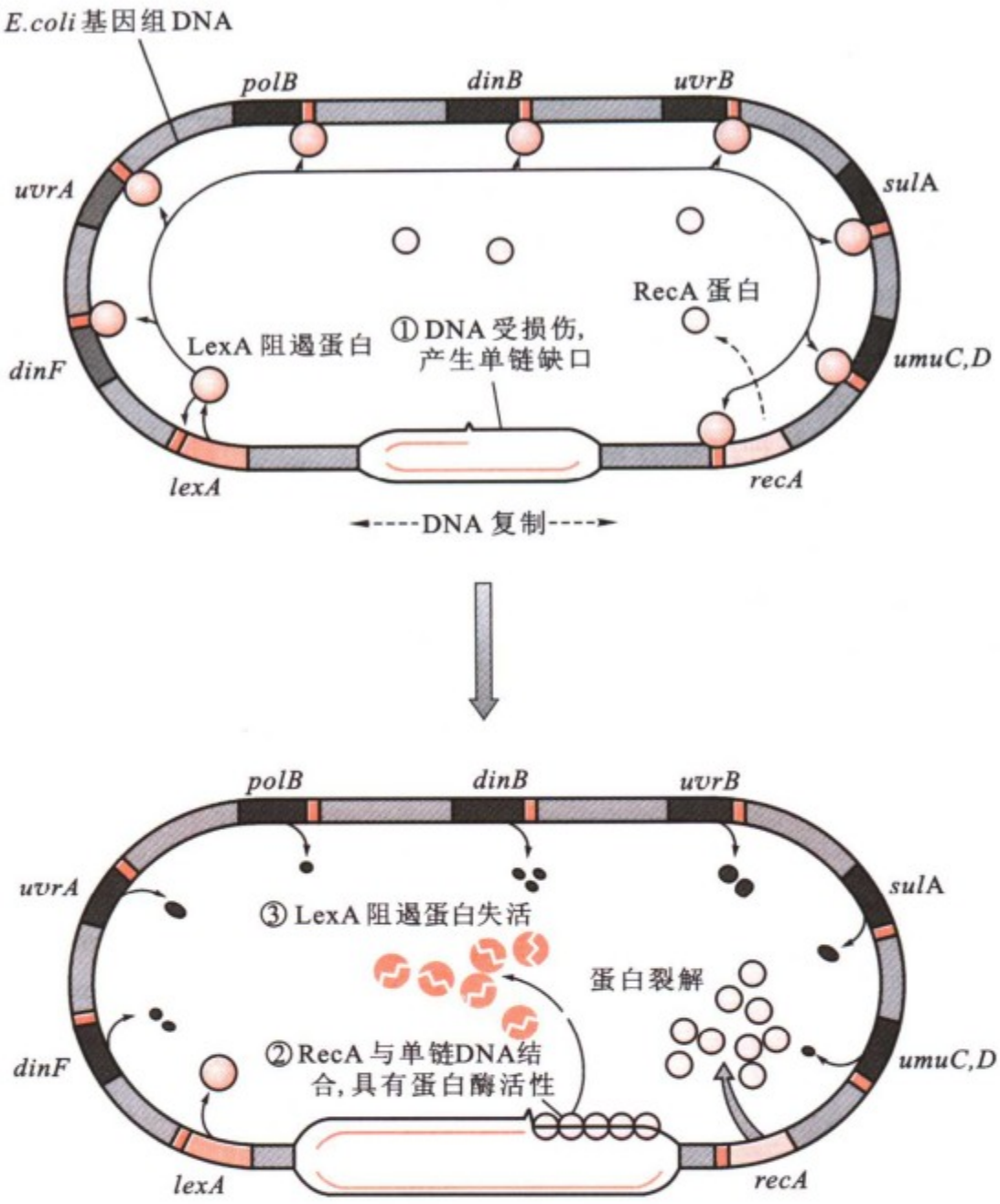



图 7-32 大肠杆菌中 LexA 蛋白阻遏的 SOS DNA 损伤修复系统操纵子的表达调控模式

导致 SOS 体系(包括 *recA* 基因)高效表达,DNA 得到修复(图 7-32)。只要有活化信号存在，该操纵子就一直处于活性状态。当修复完成后，活化信号消失，RecA 蛋白又回到非蛋白水解酶的形式,LexA 蛋白才逐步积累起来，并重新建立阻遏作用。一旦 RecA 蛋白的合成停止,细菌开始生长并进行分裂,RecA 又逐渐地稀释到原先的本底水平。

### 7.4.4 二组分调控系统和信号转导

细胞生活在不断变化的环境中,温度、pH 和渗透压变化,氧分压变化,营养的变化及细胞浓度的变化等都要求细菌必须随时作出相应的反应以求得生存。前面讨论的诱导和阻遏转录调控系统均是通过环境中的小分子效应物(诱导物或阻遏物)直接与调节蛋白结合进行转录调控,但是在较多情况下,外部信号并

不直接传递给调节蛋白，而是首先通过传感器(sensor)收信号，然后以不同方式传到调节部位，这个过程就是信号转导(signal transduction)。目前已知的最简单的细胞信号系统称为二组分系统(two-component systems)，由两种不同的蛋白质组成：即位于细胞质膜上的传感蛋白(sensor protein)，因该蛋白质具有激酶活性，所以又称传感激酶及位于细胞质中的应答调节蛋白(response regulator protein)。表 7-4 列出了大肠杆菌中存在的各种二组分调控系统。传感激酶常在与膜外环境的信号反应过程中被磷酸化，再将其磷酸基团转移到应答调节蛋白上，磷酸化的应答调节蛋白即成为阻遏或激活蛋白，通过对操纵子的阻遏或激活作用调控下游基因表达(图 7-33)。

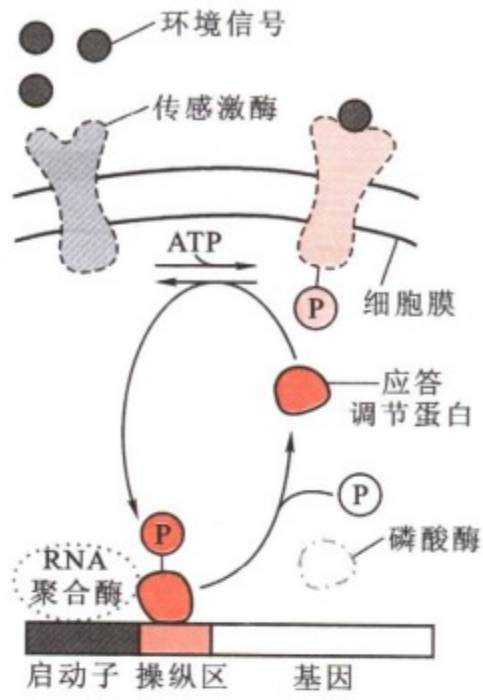



图 7-33 细菌中参与信号转导的二组分调控系统

表 7-4 大肠杆菌中存在的二组分调控系统

| 刺激信号/功能 | 传感蛋白 | 应答调节蛋白 |
|---|---|---|
| 氧气缺乏 | ArcB | ArcA |
| 渗透压，包被蛋白 | EnvZ | OmpR |
| 渗透压，钾离子运输 | KdpD | KdpE |
| 磷酸盐清除 | PhoR | PhoB |
| 氮代谢 | NtrB | NtrC |
| 硝酸盐呼吸作用 | NarX | NarL |
| 硝酸盐和亚硝酸盐呼吸作用 | NarQ | NarP |

## 7.4.5 多启动子调控的操纵子

1. rRNA 操纵子

大肠杆菌 rRNA 操纵子(*rrnE*)上有两个启动子 $P_1$ 和 $P_2$(图 7-34)。在对数生长期细菌中，$P_1$ 起始的转录产物比 $P_2$ 起始的产物多 3~5 倍，所以 $P_1$ 是强启动子。但是当细菌处于紧急状态时，如氨基酸饥饿条件下，细胞中 ppGpp 浓度增加，$P_1$ 的作用被抑制。因为 rRNA 是细胞中蛋白质合成机器——核糖体的重要组成部分，不能完全停止供应，由 $P_2$ 起始的 *rrnE* 基因转录就显得越来越重要。在贫瘠的培养基中，细胞增殖缓慢时，$P_2$ 是合成 rRNA 的主要启动子。

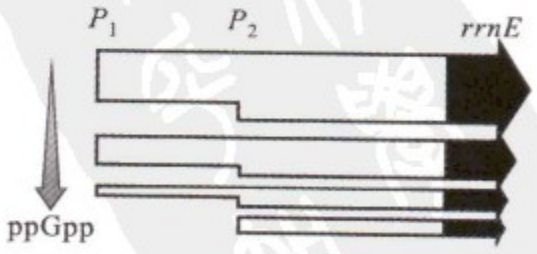



图 7-34 大肠杆菌 *rrnE* 上有两个启动子($P_1$ 和 $P_2$) 随着 ppGpp 浓度增加，$P_1$ 逐渐关闭，而 $P_2$ 活性不变。

2. 核糖体蛋白 SI 操纵子

与 *rrnE* 操纵子相类似的有核糖体蛋白SI

操纵子(*rpsA*),它也受应急反应调节。*rpsA* 有4个启动子,可能是大肠杆菌中启动子最多的操纵子。$P_1$、$P_2$ 是强启动子,平时主要依靠它们来启动基因的表达,合成 SI 蛋白。$P_3$、$P_4$ 是弱启动子,只有在紧急情况下,$P_1$、$P_2$ 启动子受 ppGpp 的抑制,由 $P_3$、$P_4$ 起始合成的 SI 蛋白维持了生命的最低需要。

3. DnaQ 蛋白操纵子

DnaQ 蛋白是 DNA 聚合酶全酶的亚基之一,其主要功能是校正 DNA 复制中可能出现的错误。已知这一操纵子受 RNA 聚合酶活性的调节,并拥有两个启动子(图 7-35)。在 RNA 聚合酶活性较低时,操纵子的转录由弱启动子 $P_2$ 控制;而 RNA 聚合酶活性较高时,就开始利用强启动子 $P_1$。因为细胞内 RNA 聚合酶活性与细胞的增殖速度有关,也就是说,在细胞染色体复制比较缓慢时,RNA 聚合酶活性往往较低,此时 DnaQ 蛋白的合成靠弱启动子 $P_2$ 来维持。当细胞增殖速度加快,RNA 聚合酶活性升高时,$P_1$ 就被激活,DnaQ 蛋白的合成就大大增加。

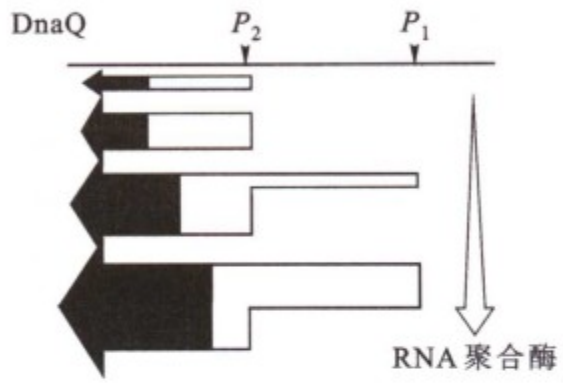



图 7-35 DnaQ 蛋白操纵子表达活性受 RNA 聚合酶调节

总之,操纵子中有不同的启动子,它们有不同的强度,其启动作用又受不同因子的调控。许多因素相互作用,才使基因表达更加有效,更加协调。在不同的生活环境中,不同的启动子精密地调节基因的表达量,这对维持细菌的生存起着非常重要的作用。

## 7.5 固氮基因调控

氮是所有生物的基本组成成分,蛋白质、核酸及许多小分子代谢产物都是含氮化合物。没有氮,就没有生命。虽然地球大气层的 80%是氮气,但是有能力直接利用这种氮源的却仅限于几类原核生物。某些豆科植物通过与根瘤菌属的固氮菌建立共生体系,也具有直接利用气态氮的本领。科学家估计,地球上的动植物共含氮约 $1.5\times10^{10}$ t,而每年通过硝化细菌以气态氮的形式释放到大气中的氮就有 $2\times10^8$~$5\times10^8$ t,全世界氮肥厂每年生产的化肥仅含氮约 $10^8$ t,所以,如果没有生物固氮,生命很快就不复存在了。

幸运的是,固氮菌从事着与硝化、脱硝化完全相反的工作,它们将气态氮转化成能被宿主植物利用的氨。根瘤则是宿主植物为适应共生固氮所产生的一种特化的器官,便于容纳类菌体并为其提供合适的环境和养分,维持氮的固定。地球上每年约有 $2\times10^8$~$5\times10^8$ t 氮通过这一途径被固定,生物固氮对地球氮代谢的贡献比所有化肥厂加起来还大。我们将在这一节里详细研究调控这一共生体系的各种因素。

### 7.5.1 根瘤的产生

根瘤菌在豆科植物根际的生长以及与宿主植物根毛幼嫩部位的接触是感

染发生的第一步。通常情况下，根瘤菌特定小种的宿主范围都是很狭窄的，如 *Rhizobium trifolii* 主要感染三叶草，*R.phaseoli* 主要感染菜豆，*R.Leguminosarum* 主要感染豌豆(表 7-5)。当然，每个小种中可能有些株系能感染宿主以外的其他豆科植物。

表 7-5 感染各种植物的不同根瘤菌

| 主要类群 | 根瘤菌名 | 宿主植物 |
|---|---|---|
| 苜蓿类 | *R. meliloti* | 苜蓿 |
| 三叶草类 | *R. trifolii* | 三叶草 |
| 豌豆类 | *R. leguminosarum* | 豌豆、某些菜豆及扁豆 |
| 菜豆类 | *R. phaseoli* | 菜豆 |
| 羽扇豆类 | *B. lupine* | 羽扇豆 |
| 大豆类 | *B. japonicum* | 大豆 |
| 豇豆类 | *Cowpea rhizobia* | 花生、豇豆、金合欢属 |

根瘤菌细胞以极性的方式与根毛接触，并附着到植物细胞上。这种接触的专一性是由植物凝血素和细菌细胞壁之间的相互作用所决定的。植物凝血素是由植物细胞所产生的一种糖蛋白，它对糖基有很高的亲和力与专一性，现已鉴定出根瘤菌细胞壁上的一种脂多糖组分很可能就是植物凝血素的受体。有实验表明，根瘤菌能产生一定量的寡聚糖来刺激植物凝血素的分泌，由基因型所决定的根瘤菌宿主范围，可能受到植物基因组的影响，因为特定的植物凝血素只能识别特定的细菌表面受体。

## 7.5.2 固氮酶

20 世纪 60 年代初期，有科学家将晾干的厌氧微生物芽胞梭菌细胞溶于稀盐溶液中，离心后取上清，加入 $^{15}N_2$，一定时间后发现反应体系中产生了 $^{15}NH_3$。进一步研究发现，无细胞提取液中的固氮酶活性会很快地、完全地被空气所破坏。

$$N_2+8H^++32ATP \longrightarrow 2NH_3+H_2+32ADP+32P_i$$

现已查明，固氮酶所催化的这一反应是通过两个步骤来完成的。组分Ⅰ是一个多聚蛋白，由两个相对分子质量各为 $5.0\times10^4$ 的 α 亚基和两个相对分子质量各为 $6.0\times10^4$ 的 β 亚基所组成，α 和 β 亚基分别由 *nifD* 和 *nifK* 基因编码。由于组分Ⅰ含有 4 个[4Fe-4S]连接桥和两个铁-钼辅助因子(FeMoCo)，所以常被称为铁钼蛋白 (FeMo-protein)。组分Ⅱ由两个完全相同的相对分子质量各为 $3.0\times10^4$ 的亚基所组成，由 *nifH* 基因编码，常被称为铁蛋白(Fe-protein)。组分Ⅱ只有一个[4Fe-4S]连接桥，位于两个亚基之间，一次只可送出一个电子，是生物固氮反应中的“瓶颈”。

除了上述两个主要组分之外，有证据说明 FeMoCo 在固氮反应中也起着举足轻重的作用。FeMoCo 可能位于酶-底结合活性中心附近，含有 1 mol Mo 与 6 mol Fe、8 mol S，能被从铁钼蛋白中分离出来。分离纯化的 FeMoCo 能被用来

活化铁钼蛋白。有肺炎克氏杆菌 *nifB*、*nifN*、*nifE* 突变体，不能产生功能型 FeMoCo，所以其铁钼蛋白无固氮活性。若加入外源 FeMoCo，铁钼蛋白的活性就得到恢复。有报道认为，铁钼蛋白 α 亚基上的第 275 位半胱氨酸与 FeMoCo 的结合是组分 I 正常发挥生理功能所必需的。进一步的研究还发现，铁钼蛋白的活性中心可能在该亚基第 191 位谷氨酰胺和第 195 位组氨酸附近，因为若分别用赖氨酸取代谷氨酰胺 $^{191}$，用天冬酰胺取代组氨酸 $^{195}$，铁钼蛋白的固氮活性明显下降(表 7-6)。

表 7-6 几种不同突变体中的固氮酶活性比较

| 株系 | 突变 | 核苷酸序列 | 铁钼蛋白相对活性 | 铁蛋白相对活性 |
|---|---|---|---|---|
| 野生型 | — | CAGTCCCTGGCCACCACATC | 40.6 | 18.7 |
| DJ357 | ΔnifDK::Kmr | — | 0 | 48.5 |
| DJ178 | α-His195→Asn | CAGTCCCTGGCAACCACATC | 3.7 | 26.3 |
| DJ255 | α-Gln191→Lys | CGCGGCGTTTCCAAGTCCCTG | 1.4 | 39.5 |

在还原氮的同时，固氮酶还有一个次要的活性，它能把氢质子还原成分子氢。这一反应不利于提高固氮酶活性，因为它既消耗了能量又产生了能竞争性抑制固氮活性的氢分子。根瘤菌中有些株系($Hup^+$)能通过氢化酶的作用催化氢和氧的结合，从而利用反应中释放的氢。这一反应使得在质子还原过程中丢失的某些 ATP 再生，还有助于除去固氮酶附近的分子氧。

### 7.5.3 与固氮有关的基因及其调控

通过对许多快速生长的根瘤菌株系，如 *R.trifolii*、*R.leguminosarum*、*R.phaseoli* 和 *R.meliloti* 的研究，发现与形成固氮根瘤有关的许多基因都位于染色体外的大质粒上。有些根瘤菌的突变体不能使植物形成根瘤($nod^-$)或者不能使植物固氮($fix^-$)，原因就在于这些质粒的特定部位发生了缺失突变。在较高温度条件下(>33℃)，培养野生型 *R.leguminosarum*($nod^+fix^+$)，使质粒 DNA 丢失，也使得这些细菌也丧失了与豌豆共生根瘤的能力。可以通过转化或转导的方法使根瘤菌的质粒在不同株系之间转移，当质粒 DNA 从野生型 $nod^+fix^+$ 转移到缺少这些功能的突变株系中，就能使这些突变体获得形成根瘤、进行生物固氮的本领。类似的实验还表明，根瘤菌的宿主范围也是由质粒决定的。例如，野生 *R.phaseoli* 只能使菜豆形成根瘤，但若消除其本身所带的质粒后导入 *R.leguminosarum* 的质粒，该重组 *R.phaseoli* 就能使豌豆形成根瘤。在 $Hup^+$ 株系中，编码氢化酶系统的基因也位于质粒 DNA 上。

1. *nifAL* 基因座对整个固氮酶系统基因转录活性的调节

实验证明，*nifAL* 基因控制了整个固氮酶系统的表达和活性。如图 7-36 所示，若突变 *nifA* 基因，那么固氮酶和其他所有已知的 *nif* 基因产物都会消失。若再将野生型 *nifA* 基因用质粒 DNA 的形式导入上述突变株中，*nif* 操纵子的功能就得到恢复。事实上，*nifA* 和 *nifL* 基因产物分别是 *nif* 操纵子(*nifAL* 本身除外)

的正和负调控因子。研究发现,当细胞内没有 nifL 蛋白质时,*nifA* 基因产物足以激活其他 *nif* 基因的表达,甚至在有 $NH_3$ 和 $O_2$ 存在时也如此。*nif* 野生型克氏肺炎杆菌 UNF928 中固氮酶的活性完全被 10 mmol/L $NH_4^+$ 所抑制,而 UNF714 由于带有突变型 *nifAL* 基因,所以固氮酶活性为零。若将带有 *nifA* 质粒 DNA 正向表达载体 pMC73A 导入 UNF714 中,不但能完全恢复野生型固氮酶活性,而且在 10 mmol/L $NH_4^+$ 中仍有 15%~30%的活性。很显然,*nifA* 基因对 UNF714 固氮酶活性的影响是通过反式作用因子来实现的,因为导入带有反义 *nifA* 基因的载体 pMC73B 没有诱发 UNF714 固氮基因转录(表 7-7)。

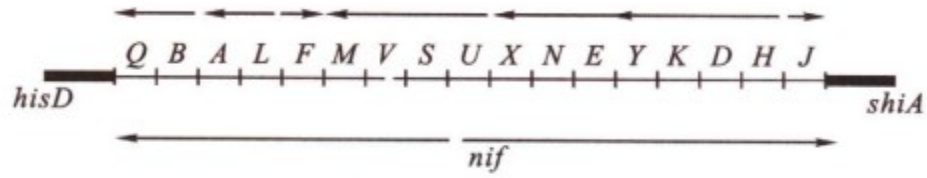



图 7-36 *nif* 操纵子结构图

表 7-7 克氏肺炎杆菌固氮酶活性与 *nifA* 质粒的关系

| 株 系 | 固氮酶活性/相对单位 | |
|---|---|---|
| | $-NH_4^+$ | $+NH_4^+$ |
| UNF928 | 100 | 0.005 |
| UNF714 | 0 | 0 |
| UNF714 + pMC73A | 230 | 27 |
| UNF714 + pMC73B | 0.1 | 0.1 |
| UNF1780(glnA100→nif) | 0.05 | 0.075 |
| UNF1780 + pMC73A | 174 | 80 |
| UNF1780 + pMC73B | 0.005 | 0.005 |

研究发现,nifA 蛋白质对 *nifH* 启动子的转录激活作用与 $NH_4^+$ 及 $O_2$ 浓度无关(表 7-8)。克氏肺炎杆菌固氮基因突变株 UNF921 由于缺乏 *nifAL* 基因产物,所以连接在 *nifH* 启动子后的 β-半乳糖苷酶活性不论有无 $NH_4^+$ 存在都几乎为零。导入带有功能型 *nifA* 基因的质粒 pMC71A 后,*nifH* 启动子所介导的 β-半乳糖苷酶活性达到 2 000 U 左右,受 nifA 蛋白激活的 *nifH* 基因表达与 $NH_4^+$ 的存在与否无关。若用上述 *nifH*::*lacZ* 融合基因构建多拷贝小型质粒 pJB31 并导入克氏肺炎杆菌中,发现它能在 *nif* 野生型株系 UNF931 中得到正常表达,并受到 $NH_4^+$ 浓度的严格控制。在带有 pJB31 质粒 DNA 的大肠杆菌 MC1061 细胞中,*nifH* 启动子无活性,所以不表达 β-半乳糖苷酶。有意思的是,若同时导入 pMC71A 质粒 DNA,*nifH* 启动子立即被激活,而且这一活性与 $NH_4^+$ 的存在基本无关。带有 pMC71A 质粒 DNA 的 UNF921 株系中 *nifH* 启动子指导的 β-半乳糖苷酶活性不受分子氧的抑制,而带有 pRD1 质粒(功能型 *nifAL* 基因)的菌株同时受到分子氧和 $NH_4^+$ 浓度的负调控,充分说明 *nifA* 基因产物在调控其他 *nif* 基因表达时的重要性。

表 7-8 *nifH* 启动子指导的 β-半乳糖苷酶基因表达与 nifA 基因产物的关系

| 细菌株系 | β-半乳糖苷酶活性/相对单位 | | | |
|---|---|---|---|---|
| | $-NH_4^+$ | | $+NH_4^+$ | |
| UNF921(lacZ:nifH) | 3 | | 3 | |
| UNF921 + pMC71A | 1 953 | | 2 267 | |
| UNF931 + pJB31 | 2 200 | | 2 | |
| MC1061 + pJB31 | 3 | | 2 | |
| MC1061 + pJB31 + pMC71A | 6 027 | | 4 246 | |
| | $+O_2$ | $-O_2$ | $+O_2$ | $-O_2$ |
| UNF921 | 3 | 6 | 2 | 3 |
| UNF921 + pMC71A | 2 250 | 2 160 | 1 570 | 3 000 |
| UNF921 + pRD1 | 22 | 2 200 | 8 | 6 |

根瘤内或体外培养的根瘤菌固氮酶活性一般受 $NH_3$ 含量和 $O_2$ 浓度两大环境因子的影响。固氮酶活性与培养基中 $NH_3$ 含量呈负相关,当培养基中 $NH_3$ 的量高于 70~100 μg/mL(4 mmol/L)时,固氮酶活性完全消失。那么,$NH_3$ 究竟是怎样决定 *nifAL* 基因表达与否呢?

2. *ntrA*、*ntrB* 及 *ntrC* 共同调控 *nifA* 基因活性

人们早就知道,有些 *nif* 突变体的突变位点在已知的 *nif* 区以外(图 7-37)。有两组这样的突变体位于质粒上 90 min *glnA* 基因附近(但不是 *glnA* 本身),分别称为 *ntrB* 和 *ntrC*(ntr,nitrogen regulation,氮调控),另有一组位于质粒上 70 min 处,称为 *ntrA*。*nifA* 基因产物只有在 *ntrA*、*ntrC* 基因产物的共同作用下才能正常工作,激活 *nif* 操纵子群,而 *ntrC* 基因(不是 *ntrA* 基因)的活性是直接受$NH_3$影响的,$NH_3$ 浓度较高时,*ntrC* 基因产物没有转录活性。看起来,*ntr* 系统像一个粗调开关,而 *nifAL* 系统则像一个微调开关,两者共同作用,控制了 *nif* 基因的表达。

分子生物学研究发现,由 *glnD* 编码的尿苷酸转移酶负责把 UMP 加到受体蛋白上。当细胞质内谷氨酰胺:α-酮戊二酸的比值较低时,UMP 就在 *glnD* 基因产物的作用下被转移到 GlnB 蛋白上,反之则从 GlnB 蛋白上脱去 UMP。不带有UMP 的 GlnB 蛋白能与 *ntrB* 产物相结合

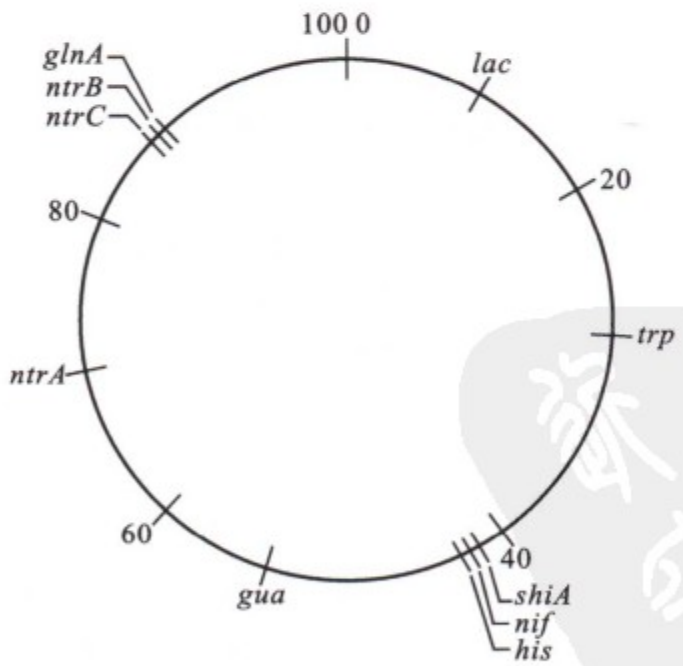



图 7-37 与固氮有关的基因在克氏肺炎杆染色体上的定位 *lac*:乳糖操纵子;*trp*:色氨酸合成酶系统;*shiA*:苯草酸合成途径酶系统;*his*:组氨酸合成途径;*gua*:鸟嘌呤合成途径;*glnA*:谷氨酰胺合酶。

并导致 *ntrC* 基因产物去磷酸化,丧失启动 *nifAL* 基因转录的功能。当细胞内游离氨含量较低时,GlnB 蛋白被尿苷化并与 NtrB 蛋白分离,使后者蛋白激酶的功能得到发挥,将 NtrC 蛋白磷酸化而激活 *nif* 操纵子系统(图 7-38)。研究表明,*ntrA* 基因编码的 δ 因子也是激活 *nif* 系统所必需的。

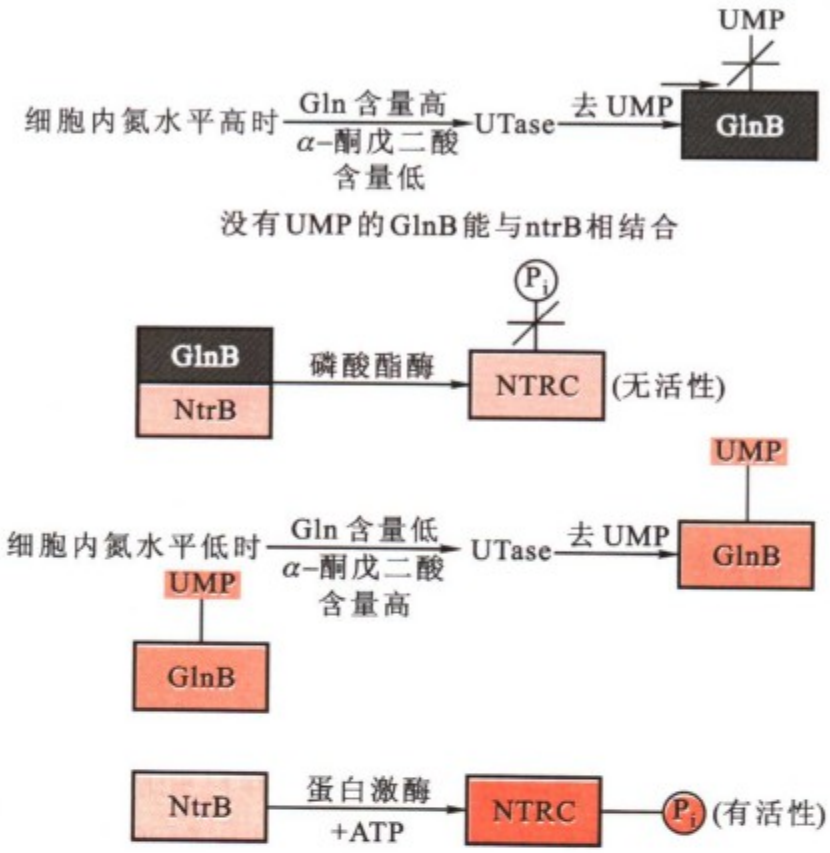



图 7-38 细胞内氮水平对 *nif* 基因转录活性的调控

3. 豆血红蛋白及根瘤素基因的调控

豆血红蛋白(leghemoglobin,Lb)在体内的主要功能是调节根瘤细胞氧含量,它是根瘤细胞中植物基因组编码的最主要的蛋白质,其血红素辅基可能是由细菌编码的。脱辅基后的 Lb 蛋白相对分子质量为 $1.56\times10^4$~$1.59\times10^4$。已经发现存在 4 种不同的 Lb 异构体(Lba、Lbc1、Lbc2 和 Lbc3),它们在氨基酸序列上虽然只有微小差异,但仍然由不同的基因编码。分析 *Lba* 和 *Lbc* 基因的核苷酸序列表明,它们的读码框中插入了 3 个长短不一的内含子,因此这些基因的总长度不一样。氨基酸序列分析证实,Lb 与动物肌红蛋白存在同源性。

除了 Lb 以外,至少还有 20 种其他多肽是由植物基因组编码但只在根瘤菌诱导的细胞中得到表达,这些蛋白质(包括 Lb)统称根瘤素(nodulin)。科学家通过被 *B.juponicum* 感染的大豆根瘤细胞对根瘤素进行了详细的研究,并根据其生理功能将其分为 3 类:第一类是根瘤结构蛋白;第二类是负责将类菌体固定的氮素同化并转移到植物体内的酶系统;第三类是维持类菌体功能所必需的蛋白质。Lb 就属于第三类根瘤素,它在根瘤固氮体系中起着不容忽视的作用,因为细胞内氧浓度一般都在数百微摩尔左右,而固氮酶只能在几十纳摩尔 $O_2$ 浓度下才能正常工作,这对矛盾就是依靠 Lb 得到解决的。如果将细胞内 $O_2$ 浓度降低到几十纳摩尔以下,呼吸系统的主要酶类早已停止了反应,根本不可能为

细胞的生命活动提供能量。Lb 对 $O_2$ 有极高的亲和力，它与大量自由 $O_2$ 相结合，既大大降低了自由 $O_2$ 浓度，又能将分子氧直接输送到呼吸链，真可谓一石二鸟，天衣无缝(表 7-9)。

表 7-9 大豆根瘤细胞内氧浓度变化分析

| | 类菌体周膜空间 | 含类菌体的宿主细胞质 |
|---|---|---|
| Lb 含量 | 300 μmol/L | 3 000 μmol/L |
| Lb 氧化程度 | 0.2 | 0.2 |
| 氧吸附量 | 60 μmol/L | 600 μmol/L |
| 自由氧浓度 | 11 nmol/L | 11 nmol/L |
| 吸附氧/自由氧 | 5 500 | 55 000 |

曾有人用大豆根瘤特异性 cDNA 进行选择性杂交，并利用与多核糖体相结合的多聚腺苷化 RNA 进行体外翻译，发现每一个含类菌体的根瘤细胞内可有 600 000 个与核糖体相结合的 mRNA 分子，其中大约有 90 000 个(15%)是 Lb mRNA 分子。总之，共生固氮是一个很有意思的体系。人类虽然不一定能在不久的将来使禾谷类植物固氮，但对豆科植物与固氮菌的相互识别及固氮基因调控因子的研究，却一定可以更加丰富我们对植物和微生物界的认识。

## 7.6 转录水平上的其他调控方式

转录过程涉及转录机器附着于 DNA，识别启动子序列，起始 RNA 的合成，延伸和终止。转录的任何一步都受到调控，其中某些步骤是主要的调控点，比如 RNA 转录的起始相对延伸和终止来说。自然选择使得生物的调控策略达到最优化。转录过程的调节是生物基因表达调节最有效的方式。

### 7.6.1 σ 因子的调节作用

不同的 σ 因子可以独立地起作用，但是，为了保证细胞准确响应不同的环境信号变化，σ 因子之间常常交互作用构成网络调控模式，使得原核基因的表达稳定而平衡。σ 因子本身的活性受到蛋白水解酶的调控，也能被同源的抗 σ 因子失活。这些抗 σ 因子能够与特定的 σ 因子结合，阻止它们与 RNA 聚合酶的组装。

营养物缺乏时，革兰阳性菌 *Bacillus* 会形成孢子来渡过困难时期。其孢子的形成需要 4 种不同的 σ 因子。$\sigma^E$ 和 $\sigma^K$ 存在于母细胞中，一旦开始形成孢子就产生 $\sigma^F$ 和 $\sigma^G$。$\sigma^E$ 和 $\sigma^K$ 先以非活性的前体形式被合成，经过特定的蛋白酶作用转变成活性形式。$\sigma^F$ 也是以非活性形式(与抗 σ 因子 SpoⅡAB 结合)存在于胞子中，环境刺激导致 SpoⅡ AA 抗-抗 σ 因子去磷酸化，并特异性地与抗 σ 因子SpoⅡAB 结合，释放出有活性的$\sigma^F$(图 7-39)。活性$\sigma^F$ 促使早期孢子形成相关基因，包括$\sigma^G$ 和需要进入母细胞中降解前体 $\sigma^E$ 蛋白酶基因的转录。活性$\sigma^G$ 主

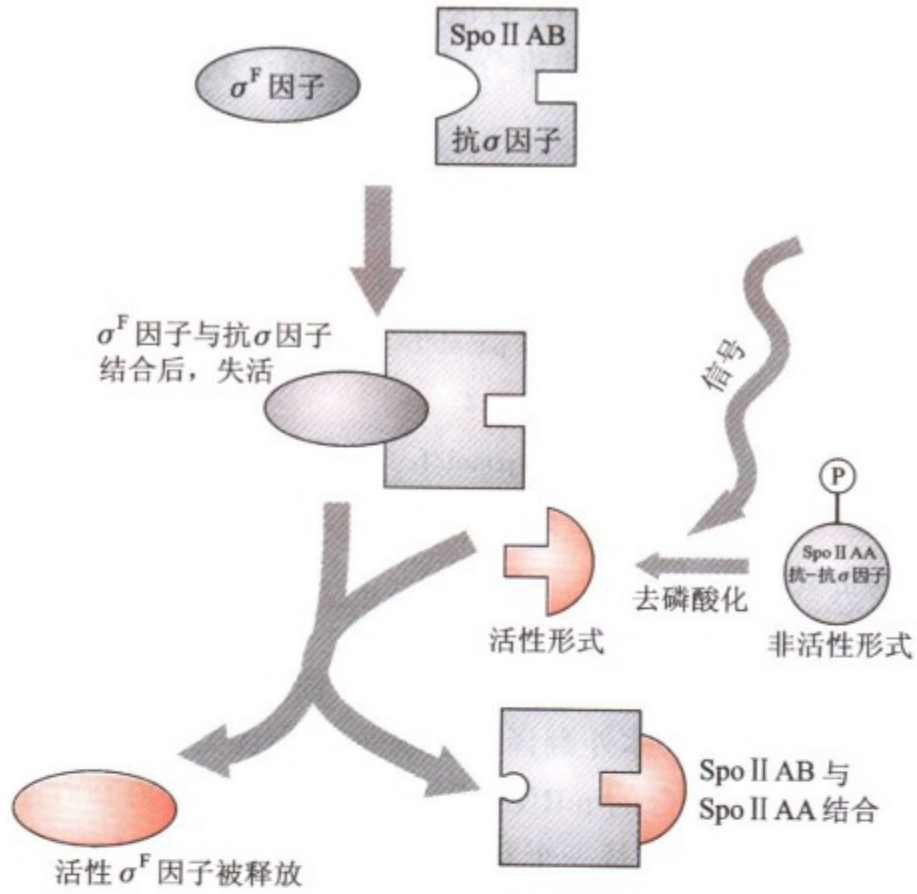



图 7-39 $\sigma^F$的两种存在形式

要激活后期孢子形成相关基因和需要进入母细胞中降解前体$\sigma^K$的蛋白酶基因的转录。

### 7.6.2 组蛋白类似蛋白的调节作用

细菌中存在一些非特异性的DNA结合蛋白，用来维持DNA的高级结构，被称为组蛋白类似蛋白(histone-like proteins)。细菌中的H-NS蛋白,就以非特异性的方式结合DNA,维持其高级结构。H-NS包含2个结构域,一个DNA结合结构域和一个蛋白-蛋白相互作用结构域。H-NS先结合到DNA上,然后通过蛋白-蛋白相互作用形成四聚体或者多聚体,帮助维持DNA的高级结构。另外,H-NS与大肠杆菌基因组上分散的大量基因的调控区有较高的亲和性，这些基因大都与环境条件的变化有关。H-NS非特异性结合在这些DNA上,抑制这些基因的转录。这些基因的转录激活需要特定的转录因子参与。

### 7.6.3 转录调控因子的作用

能够与基因的启动子区相结合，对基因的转录起激活或抑制作用的DNA结合蛋白被称为转录调控因子。大肠杆菌基因组中有300多个基因编码这样的蛋白质，它们大多数是序列特异性的DNA结合蛋白，能够与特定的启动子结合。有些能够调控大量基因的表达,而有些仅调控一两个基因的表达。转录调控因子CRP、FNR、IHF、Fis、ArcA、NarL和Lrp调控了50%基因的表达，而约有60个转录因子仅能特异性结合一两个启动子。有些转录因子对某个基因起激活作

用,却对另一个基因起抑制作用。有些转录因子对同一基因也能发挥两种不同的作用,如 AraC 蛋白在结合阿拉伯糖前后就分别起着抑制和激活阿拉伯糖操纵子基因转录的作用。许多基因的启动子区有多个转录调控因子的结合位点,这些转录调控因子的共同作用才能使 RNA 聚合酶顺利地结合在 DNA 上起始基因转录的过程。

### 7.6.4 抗终止因子的调节作用

抗终止因子是能够在特定位点阻止转录终止的一类蛋白质。当这些蛋白质存在时,RNA 聚合酶能够越过终止子,继续转录 DNA。这种基因表达调控机制主要见于噬菌体和少数细菌中。在 RNA 聚合酶到达终止子之前与 RNA 聚合酶结合,因为在终止子上游存在抗终止作用的信号序列,只有与抗终止因子相结合的 RNA 聚合酶才能顺利通过具有茎环结构的终止子,使转录继续进行(图 7-40)。

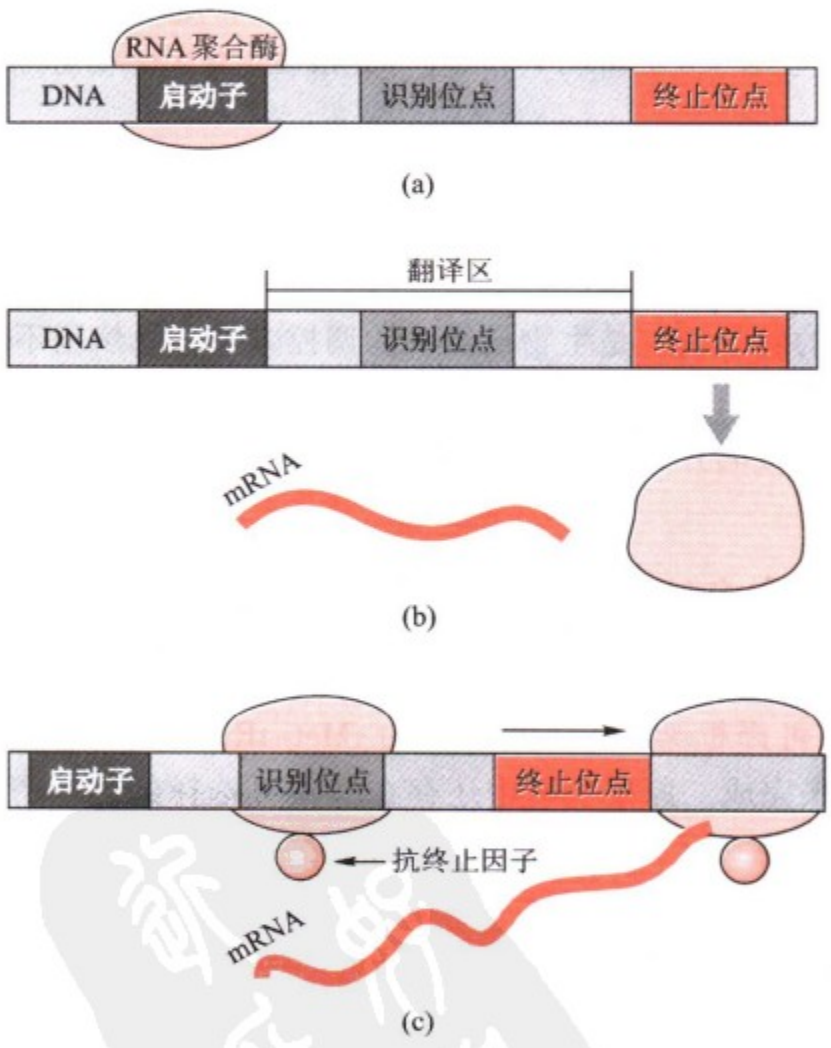



**图 7-40 抗终止子的通读作用** (a)开始转录;(b)缺少抗终止因子;(c)存在抗终止因子。

参与大肠杆菌抗终止作用的蛋白是 Nus 蛋白。转录起始不久,$\sigma$因子从 RNA 聚合酶上解离下来,NusA 蛋白就结合到了核心 RNA 聚合酶上。NusA 的结合增加了 RNA 聚合酶在终止子发夹结构处暂停的过程,从而促进抗终止作用的发生。NusA 和 $\sigma$因子不能同时结合到 RNA 聚合酶上。只要 RNA 聚合酶结合

在DNA 上,NusA 就不会从 RNA 聚合酶上解离，而 $\sigma$ 因子可以取代游离 RNA 聚合酶上的 NusA,因此,在转录起始和终止的过程中 RNA 聚合酶分别受到 $\sigma$ 因子和 NusA 的调控。大肠杆菌核糖体 RNA 编码基因 *rrn* 的转录就是抗终止转录的典型。当结合有 NusA 的 RNA 聚合酶遇到 boxA 终止序列时,NusB/S10 抗终止因子在 NusG 的帮助下与 RNA 聚合酶结合,实现抗终止(图 7-41)。

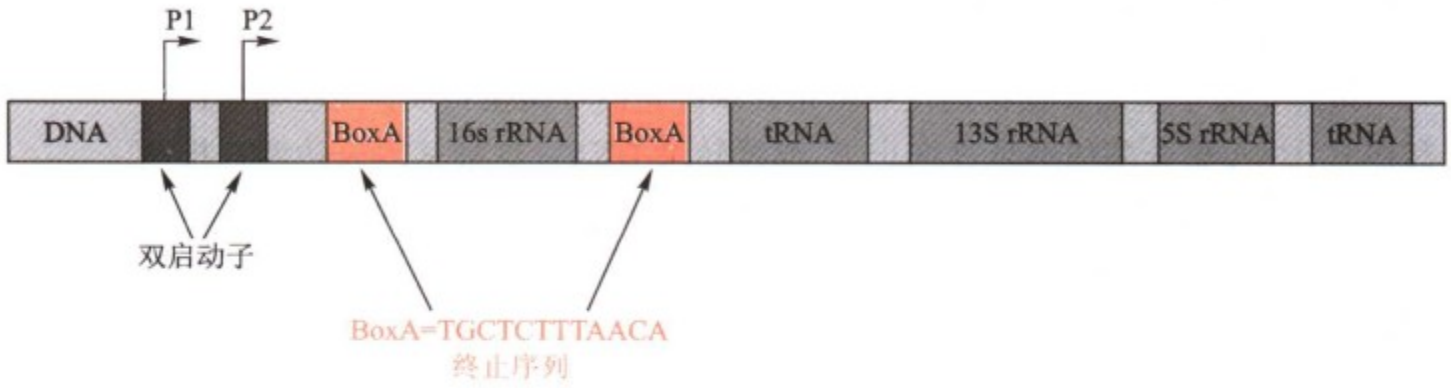



**图 7-41 抗终止子参与大肠杆菌 *rrn* 基因转录的调控** 当结合有 NusA 的 RNA 聚合酶遇到 BoxA 终止序列时,NusB/S10 抗终止因子可在 NusG 的作用下与 RNA 聚合酶结合,实现抗终止。

## 7.7 转录后调控

基因表达的转录调控是生物最经济的调控方式，既然用不着某种蛋白质，就大可不必转录了。但转录生成 mRNA 以后,再在翻译或翻译后水平进行“微调”,是对转录调控的补充,它使基因表达的调控更加适应生物本身的需求和外界条件的变化。

### 7.7.1 mRNA 自身结构元件对翻译起始的调节

原核生物的翻译要靠核糖体 30S 亚基识别 mRNA 上的起始密码子 AUG,以此决定它的可译框架,AUG 的识别由 fMet-tRNA 中含有的碱基配对信息(3′-UAC-5′)来完成。原核生物中还存在其他可选择的起始密码子,14%的大肠杆菌基因起始密码子为 GUG,3%为 UUG,另有两个基因使用 AUU。这些不常见的起始密码子与 fMet-tRNA 的配对能力较 AUG 弱,从而导致翻译效率的降低。有研究表明,当 AUG 被替换成 GUG 或 UUG 后,mRNA 的翻译效率降低了 8倍。

遗传信息翻译成多肽链起始于 mRNA 上的核糖体结合位点(ribosome binding site,RBS)，一般是起始密子 AUG 上游的包括 SD 序列在内的一段非翻译区,该序列与核糖体 16S rRNA 的 3′端互补配对,促使核糖体结合到 mRNA 上,有利于翻译的起始。RBS 的结合强度取决于 SD 序列的结构及其与起始密码 AUG 之间的距离。SD 与 AUG 之间相距一般以 4~10 个核苷酸为佳,9 个核苷酸为最佳。

此外,mRNA 的二级结构也是翻译起始调控的重要因素。因为核糖体的 30S

亚基必须与 mRNA 结合，要求 mRNA5′端有一定的空间结构。SD 序列的微小变化，往往会导致表达效率上百倍甚至上千倍的差异，这是由于核苷酸的变化改变了形成 mRNA5′端二级结构的自由能，影响了核糖体 30S 亚基与 mRNA 的结合，从而造成了蛋白质合成效率上的差异。

## 7.7.2 mRNA 稳定性对转录水平的影响

所有细胞都有一系列核酸酶，用来清除无用的 mRNA。一个典型的细菌 mRNA 半衰期为 2~3 min。mRNA 分子被降解的可能性取决于它们的二级结构。大肠杆菌有一个 CsrAB 调节系统。CsrA 为一个 RNA 结合蛋白，CsrB 是一个非编码的 RNA 分子。CsrA 可以结合到受其调控的 mRNA 上，也可以与 CsrB 的 RNA 分子相结合。每个 CsrB RNA 分子最多可结合 18 个 CsrA 蛋白。

在静止期细菌细胞内，糖以糖原的形式被储藏起来；在细胞快速生长周期，通过糖酵解途径消耗糖，这两个过程的平衡就是由 CsrAB 调节系统来完成的。CsrA 蛋白可激活糖酵解过程并抑制葡萄糖和糖原的合成。在糖原合成途径中，如果 CsrA 蛋白结合到 *glg* 基因的 mRNA 分子上，该 mRNA 分子就易于受核酸酶攻击，其降解过程加快，其作为蛋白质合成模板的功能就受到抑制(图 7-42)。

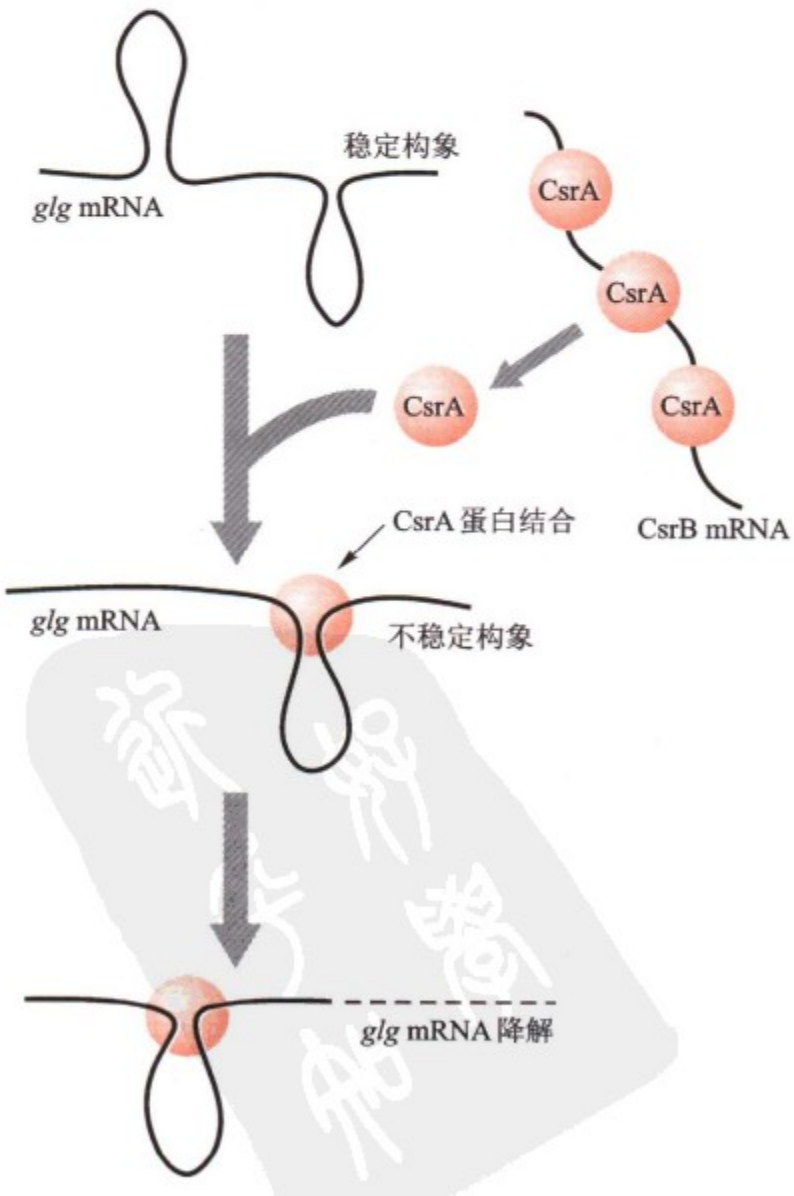



图 7-42 CsrA 蛋白调控 *glg* mRNA 的稳定性

### 7.7.3 调节蛋白的调控作用

细菌中有些 mRNA 结合蛋白可激活靶基因的翻译。大肠杆菌 BipA 蛋白就具有依赖于核糖体的 GTP 酶活性,能够激活转录调控蛋白 fis mRNA 的翻译,是 fis 蛋白质合成所必需的。

相反,mRNA 特异性抑制蛋白则通过与核糖体竞争性结合 mRNA 分子来抑制翻译的起始。大肠杆菌中核糖体蛋白就存在翻译抑制现象。*rRNA* 基因正常的转录和翻译将产生核糖体蛋白,由于这些蛋白与 rRNA 分子的亲和力较强,只要细胞中有足够的 rRNA 分子,核糖体蛋白就不会结合到自身 mRNA 分子上。当细胞中缺乏足够的 rRNA 分子时,核糖体蛋白只能结合到自身 mRNA 分子上,导致该 mRNA 的 RBS 位点被封闭, 蛋白质合成停止 (图 7–43)。这种机制保证了 rRNA 和核糖体蛋白在数量上的平衡。

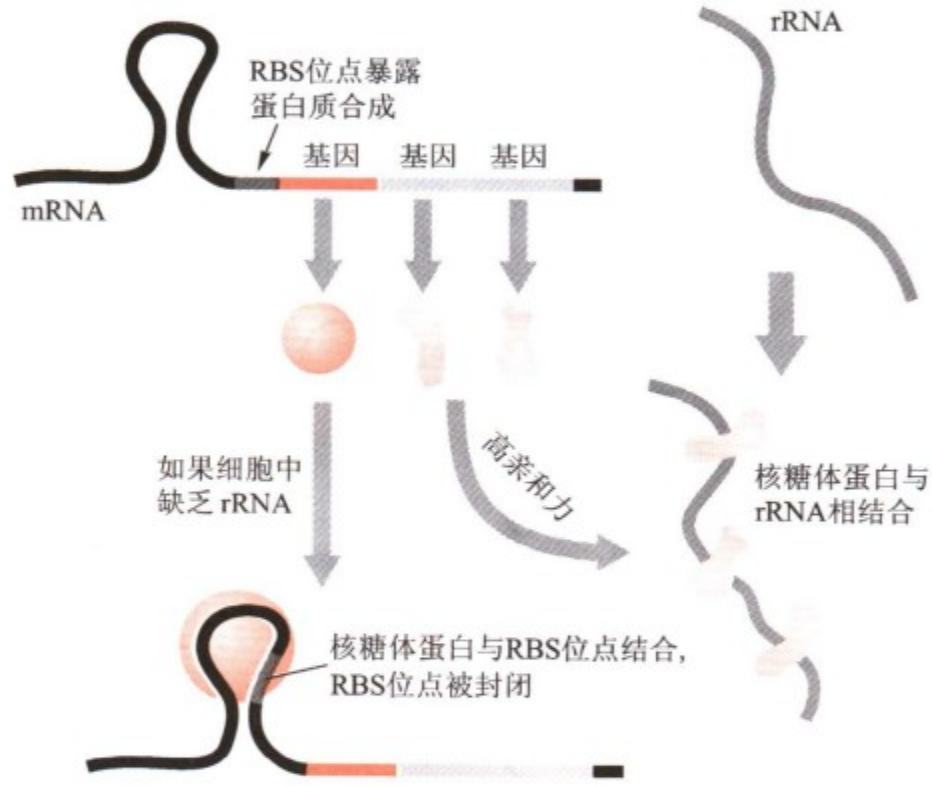



图 7–43 核糖体蛋白对自身 mRNA 翻译的抑制作用

### 7.7.4 反义 RNA 的调节作用

RNA 调节是原核生物基因表达转录后调节的另一种重要机制。细菌响应环境压力(氧化压力、渗透压、温度等)的改变,会产生一些非编码的小 RNA 分子,能与 mRNA 中的特定序列配对并改变所配对 mRNA 分子的构象, 导致翻译过程被开启或者关闭,也可能导致目标 mRNA 分子的快速降解。

细菌铁蛋白用来储存细胞中过剩的铁离子。*bfr* 基因编码细菌铁蛋白,而 anti–*bfr* 基因编码反义 RNA。培养基中铁离子浓度较低时,不需要细菌铁蛋白;当铁离子浓度升高时,就需要细菌铁蛋白行使功能。然而,细胞中无论铁离子浓度高低,*bfr* 基因都正常转录成 mRNA, 而 anti–*bfr* 基因的转录却受到能够感应

铁离子浓度变化的 Fur 蛋白的调控。细胞中铁离子过多时，Fur 蛋白作为抑制因子起作用，关掉与铁摄取有关的众多操纵子，包括 anti-*bfr* 基因的表达，使 *bfr* 基因能正常翻译出细菌铁蛋白，储存过剩的铁离子。在铁离子浓度低时，anti-*bfr* 基因转录产生大量反义 RNA，与 *bfr* 的 mRNA 配对，阻止细菌铁蛋白基因的翻译（图 7-44）。

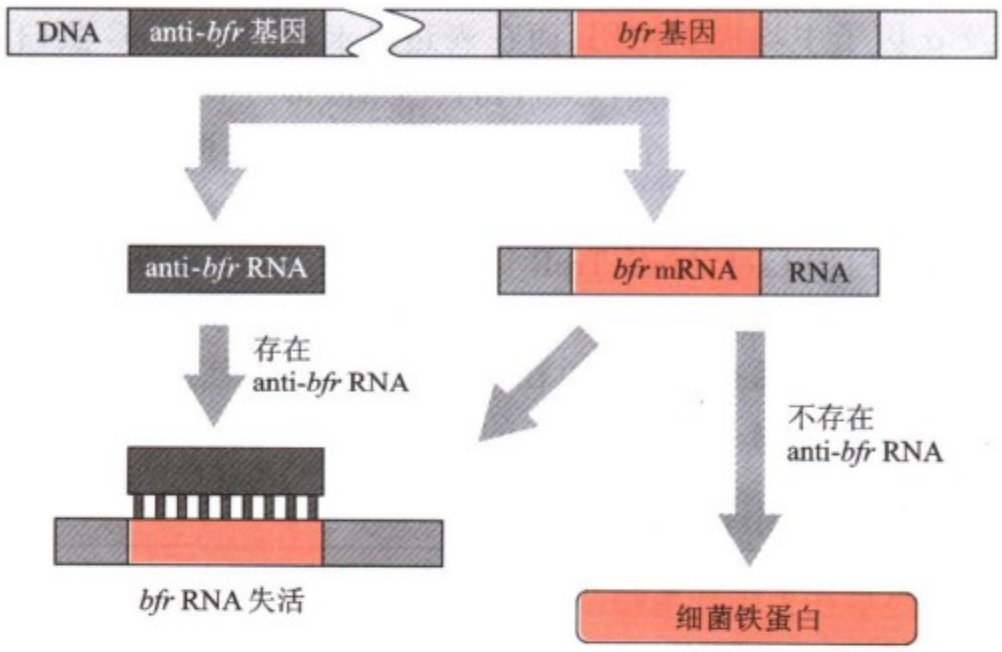



图 7-44 反义 RNA 对 *bfr* 基因翻译的抑制作用

此外，大肠杆菌中受氧化压力诱导产生的 OxyS 和受低温产生的 *DsrA* 基因也可以反义 RNA 的方式调控多个基因的翻译。OxyS 可以与 *fhlA*（编码细菌甲酸盐代谢中的转录激活蛋白）mRNA 的核糖体结合位点的部分序列配对，也可以与 *fhlA* 编码序列中的特定序列配对，阻止核糖体的结合从而抑制 *fhlA* mRNA 的翻译。*DsrA* 则能促进编码压力相关 $\sigma$ 因子 *rpoS* 基因的翻译。*DsrA* RNA 具有三个茎环，其第一个茎环结构序列与 *rpoS* mRNA 上游的前导序列互补配对，破坏了 *rpoS* mRNA 核糖体结合位点序列处的茎环结构，释放核糖体结合位点，从而提高了翻译效率。

### 7.7.5 稀有密码子对翻译的影响

大肠杆菌 DNA 复制时冈崎片段之前的 RNA 引物是由 *dnaG* 基因编码的引物酶催化合成的，细胞对这种酶的需求量不大，而引物酶过多对细胞是有害的。已知 *dnaG*、*rpoD* 及 *rpsU*（30S 核糖体上的 S21 蛋白）属于大肠杆菌基因组上的同一个操纵子，而这 3 个基因产物在数量上却大不相同，每个细胞内仅有 50 拷贝的 dnaG 蛋白，却有 2 800 拷贝的 rpoD 蛋白，更有高达 40 000 拷贝的 rpsU 蛋白。细胞通过翻译调控，解决了这个问题。

研究 *dnaG* 序列发现其中含有不少稀有密码子。分别计算大肠杆菌中 25 种结构蛋白和 *dnaG*、*rpoD* 序列中 64 种密码子的利用频率，可以看出 *dnaG* 与其他两类有明显不同（表 7-10）。很明显，稀有密码子 AUA 在高效表达的

表 7-10 几种蛋白质中异亮氨酸密码子使用频率比较

| 蛋白质 | AUU(%) | AUC(%) | AUA(%) |
|---|---|---|---|
| 结构蛋白 | 37 | 62 | 1 |
| σ 亚基 | 26 | 74 | 0 |
| DnaG 蛋白 | 36 | 32 | 32 |

结构蛋白及 σ 因子中均极少使用,而在表达要求较低的 dnaG 蛋白中使用频率就相当高。此外,UCG(Ser)、CCU(Pro)、CCC(Pro)、ACG(Thr)、CAA(Gln)、AAT(Asn)和 AGG(Arg)等 7 个密码子的使用频率在不同蛋白中也有明显差异。

许多调控蛋白如 LacI、AraC、TrpR 等在细胞内含量也很低,这些蛋白的编码基因中密码子的使用频率和 *dnaG* 相似,而明显不同于结构蛋白。科学家认为,由于细胞内对应于稀有密码子的 tRNA 较少,高频率使用这些密码子的基因翻译过程容易受阻,影响了蛋白质合成的总量。

### 7.7.6 重叠基因对翻译的影响

重叠基因最早在大肠杆菌噬菌体 ΦX174 中发现,例如,*B* 基因包含在 *A* 基因内,*E* 基因包含在 *D* 基因内,通过不同的阅读方式可得到不同的蛋白质。当时认为重叠基因的生物学意义是它可以包含更多的遗传信息,后来发现丝状RNA噬菌体、线粒体 DNA 和细菌染色体上都有重叠基因存在,暗示这一现象可能对基因表达调控有影响。现用 *trp* 操纵子中的 *trpE* 基因和 *trpD* 基因之间的翻译偶联现象来说明这个问题。

*trp* 操纵子由 5 个基因(*trpE*、*D*、*C*、*B*、*A*)组成,在正常情况下,操纵子中 5 个基因产物是等量的,但 *trpE* 突变后,其邻近的 *trpD* 产量比下游的 *trpBA* 产量要低得多。研究 *trpE* 和 *trpD* 以及 *trpB* 和 *trpA* 两对基因中核苷酸序列与翻译偶联的关系,发现 *trpE* 基因的终止密码子和 *trpD* 基因的起始密码子共用一个核苷酸。

*trpE*——苏氨酸——苯丙氨酸——终止

ACU ——UUC —— UGA uggcu

AUG —— GCU

甲硫氨酸—丙氨酸—*trpD*

由于 *trpE* 的终止密码子与 *trpD* 的起始密码重叠,*trpE* 翻译终止时核糖体立即处在起始环境中,这种重叠的密码保证了同一核糖体对两个连续基因进行翻译的顺畅。实验证明,偶联翻译可能是保证两个基因产物在数量上相等的重要手段。除了上述 *trpE* 和 *trpD* 基因之外,*trpB* 和 *trpA* 基因也存在着翻译偶联现象,因为这两个基因的产物等量存在于细胞中,而其核苷酸序列同样是重叠的。

*trpB*——谷氨酸——异亮氨酸——终止

GAA ——AUC —— UGAuggaa

AUG——GAA

甲硫氨酸——谷氨酸——*trpA*

此外,大肠杆菌 *gal* 操纵子(*galETK*)中也存在着一种基因重叠现象。*galT* 的终止密码子虽然与 *galK* 的起始密码子相隔 3 个核苷酸,*galK* 基因的 SD 序列却位于 *galT* 基因终止密码子之前,尽管这一现象与前述的偶联翻译并不完全相同,但因为核糖体结合在 mRNA 上,可以覆盖 20 个核苷酸,包括 SD 序列和 *galK* 基因的起始密码子,所以当 *galT* 翻译终止时,核糖体还没有脱落就直接与SD 序列结合,开始 *galK* 的翻译,能保证两个基因的等量翻译,也是一种翻译偶联。

### 7.7.7 翻译的阻遏

蛋白质阻遏或激活基因转录的例子已经屡见不鲜,那么,蛋白质是否也能对翻译起类似的调控作用呢? 在大肠杆菌 RNA 噬菌体 Qβ 中发现了这种现象。Qβ 噬菌体基因组包含有 3 个基因, 从 5′到 3′方向依次是与噬菌体组装和吸附有关的成熟蛋白基因 *A*, 外壳蛋白基因和 RNA 复制酶基因。当噬菌体感染细菌,RNA 进入细胞后,这条称为(+)链的 RNA 立即作为模板指导合成复制酶,并与宿主中已有的亚基结合行使复制功能。但是,Qβ(+)RNA 链上此时已有不少核糖体,它们从 5′到 3′方向进行翻译,这无疑影响了复制酶催化的从 3′到 5′方向进行的(−)链合成。克服这个矛盾的办法便是由 Qβ 复制酶作为翻译阻遏物进行调节。

体外实验证明,纯化的复制酶可以和外壳蛋白的翻译起始区结合,抑制蛋白质的合成。由于复制酶的存在,核糖体便不能与起始区结合,但已经起始的翻译仍能继续下去,直到翻译完毕,核糖体脱落,与(+)RNA 3′端结合的复制酶便开始了 RNA 的复制。这里复制酶既能与外壳蛋白的翻译起始区结合, 又能与(+)链 RNA 的 3′端结合。序列分析表明,这两个位点上都有 CUUUUAAA 序列,能形成稳定的茎–环结构,具备翻译阻遏特征。

### 7.7.8 魔斑核苷酸水平对翻译的影响

实验证明,核糖体蛋白(R 蛋白)通过反馈阻遏作用保证所有 R 蛋白合成的协同进行。那么,细胞如何保证蛋白质合成的总速度与蛋白质合成机器主要成分 rRNA 的合成速率相一致? 保证在不进行蛋白质合成时没有 RNA 的合成? 起这个调控作用的最早信号是不载有氨基酸的 tRNA。

大肠杆菌营养缺陷型($trp^-his^-$)在缺少任何一种必需氨基酸的培养基上生长时,不但蛋白质合成速度立即下降,RNA 合成速度也下降。由于色氨酸和组氨酸不是 RNA 的前体, 因此认为 RNA 合成速度下降是蛋白质合成受阻后的次级反应。研究另一个大肠杆菌突变株发现,当氨基酸供应不足时,这个突变株细胞内蛋白质合成虽然停止了,但 RNA 的合成速度却没有下降。科学上把前一种现象称为严紧控制 (基因型 $rel^+$); 后一种现象称为松散控制 (基因型 $rel^-$)。

$rel^+$和 $rel^-$菌株除上述生理现象不同之外, 缺乏氨基酸时 $rel^+$菌株能合成鸟苷四磷酸(ppGpp)和鸟苷五磷酸(pppGpp),$rel^-$菌株则不能合成鸟苷酸。因为这

两种化合物是在色谱上检出的斑点,当时称为魔斑(magic spot)。在旺盛生长的细胞中约有 65%~90% tRNA 是载有氨基酸的。当氨基酸缺乏时,不负载氨基酸的 tRNA 增多,这种不负载氨基酸的 tRNA 仍能与核糖体的 A 位结合,核糖体上不负载氨基酸的 tRNA 是细胞产生严紧控制的信号。在正常的蛋白质合成过程中,将 AA-tRNA 转运到正在延伸的多肽上需要 GTP,也许由于这一反应的停止,大量 GTP 便被用作合成魔斑核苷酸的前体。

$$GTP+ATP \longrightarrow pppGpp+AMP \longrightarrow ppGpp$$

参与这个反应的除 *relA* 基因所编码的 ATP-GTP 3′焦磷酸转移酶外,还需要翻译起始因子 EFTu 和 EFG 等。虽然 ppGpp 是多效的,但它的主要作用可能是影响RNA 聚合酶与启动子结合的专一性,从而成为细胞内严紧控制的关键。有人将 ppGpp 和 cAMP 这类物质称为警报素(alarmone)。当细胞缺乏氨基酸时产生 ppGpp,可在很大范围内做出如抑制核糖体和其他大分子合成等应急反应,活化某些氨基酸操纵子的转录表达,抑制与氨基酸转运无关的系统,活化蛋白水解酶等,以节省或开发能源,渡过难关。

## 思考题

1. 简述代谢物对基因表达调控的两种方式。
2. 什么是操纵子学说?
3. 简述乳糖操纵子的调控模型。
4. 什么是葡萄糖效应?
5. 什么是弱化作用?
6. 简述抗终止子的调控机制。
7. 简述反义 RNA 的调控机制。
8. 简述原核基因转录后调控的不同方式。

## 主要参考文献


1. Argaman L, Altuvia S. *fhlA* repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. J. Mol. Biol. 2000, 300: 1101-1112.
2. Barnard A, Wolfe A, Busby S. Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. Curr Opin Microbiol. 2004, 7: 102-108.
3. Benoff B, Yang H, Lawson C L *et al*. Structural Basis of Transcription Activation: The CAP-alpha CTD-DNA Complex. Science, 2002, 297: 1562-1566.
4. Browning D F, Busby S J. The regulation of bacterial transcription initiation. Nat. Rev. Microb. 2004, 2: 57-65.
5. Fang F C. Sigma cascades in prokaryotic regulatory networks. Proc. Natl. Acad.

Sci. U.S.A 2005,102:4933-4934.

6. Jenner L,Romby P,Rees B *et al.* Translational operator of mRNA on the ribosome: how repressor proteins exclude ribosome binding. Science,2005,308: 120-123.
7. Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene,2005,361:13-37.
8. Owens R M,Pritchard G,Skipp P *et al.* A dedicated translation factor controls the synthesis of the global regulator Fis. EMBO J. 2004,23:3375-3385.
9. Wilson C J,Zhan H,Swint-Kruse L *et al.* The lactose repressor system: paradigms for regulation,allosteric behavior and protein folding.Cell Mol. Life Sci. 2007,64: 3-16.
10. Worhunsky D J,Godek K,Litsch S *et al.* Interactions of the Non-coding RNA DsrA and RpoS mRNA with the 30 S Ribosomal Subunit.J. Biol. Chem. 2003, 278:15815-15824.
11. Yanofsky C. The different roles of tryptophan transfer RNA. in regulating trp operon expression in *E.coli* versus *B.subtilis*. Trends Genet. 2004,20:367-374.

# 8 基因的表达与调控(下)

## ——真核基因表达调控一般规律


- 8.1 真核生物的基因结构与转录活性
- 8.2 真核基因转录机器的主要组成
- 8.3 蛋白质磷酸化对基因转录的调控
- 8.4 蛋白质乙酰化对基因表达的影响
- 8.5 激素与热激蛋白对基因表达的影响
- 8.6 其他水平上的表达调控


真核生物和原核生物在基因表达调控上的巨大差别是由两者基本生活方式不同所决定的。原核生物一般为自由生活的单细胞,只要环境条件合适,养料供应充分,它们就能无限生长、分裂,因此,它们的调控系统就是要在一个特定的环境中为细胞创造高速生长的基础,或使细胞在受到损伤时,尽快得到修复。它们主要是通过转录调控,以开启或关闭某些基因的表达来适应环境条件(主要是营养水平的变化)。环境因子往往是调控的诱导物,群体中每个细胞对环境变化的反应都是直接和基本一致的。

真核生物(除酵母、藻类和原生动物等单细胞类之外)主要由多细胞组成,每个真核细胞所携带的基因数量及总基因组中蕴藏的遗传信息量都大大高于原核生物,如人类细胞单倍体基因组就包含有 $3\times10^9$bp 总 DNA,约为大肠杆菌总 DNA 的 800 倍,是噬菌体总 DNA 的 10 万倍左右!真核生物基因组 DNA 中有许多重复序列,基因内部还常插入不翻译成蛋白质的序列,都影响了真核基因的表达。真核生物的 DNA 还常与蛋白质(包括组蛋白和非组蛋白)结合,形成十分复杂的染色质结构。染色质构象的变化、染色质中蛋白质的变化及染色质对 DNA 酶敏感程度的变化等都会对基因表达产生重要影响。此外,真核生物染色质被包裹在细胞核内,基因的转录(核内)和翻译(细胞质内)被核膜隔开,核内 RNA 的合成与转运,细胞质中 RNA 的剪接和加工等无不扩大了真核生物基因调控的范围,使真核生物基因调控达到了原核生物所不可能拥有的深度和广度。对大多数真核细胞来说,基因表达调控的最明显特征是能在特定时间和特定的细胞中激活特定的基因,从而实现“预定”的、有序的、不可逆转的分化、发育过程,并

使生物的组织和器官保持正常功能。

真核生物基因调控可分为两大类，第一类是瞬时调控或称可逆性调控，它相当于原核细胞对环境条件变化所做出的反应，包括某种底物或激素水平升降时，或细胞周期不同阶段中酶活性的调节；第二类是发育调控或称不可逆调控，是真核基因调控的精髓部分，它决定了真核细胞生长、分化、发育的全部进程。根据基因调控在同一事件中发生的先后次序，又可将其分为转录水平调控(transcriptional regulation)及转录后水平调控(post-transcriptional regulation)，后者又进一步分为 RNA 加工成熟过程的调控(RNA processing)、翻译水平的调控(translational regulation)及蛋白质加工水平的调控(protein maturation and processing)等(图 8-1)。研究基因调控主要应回答 3 个问题：

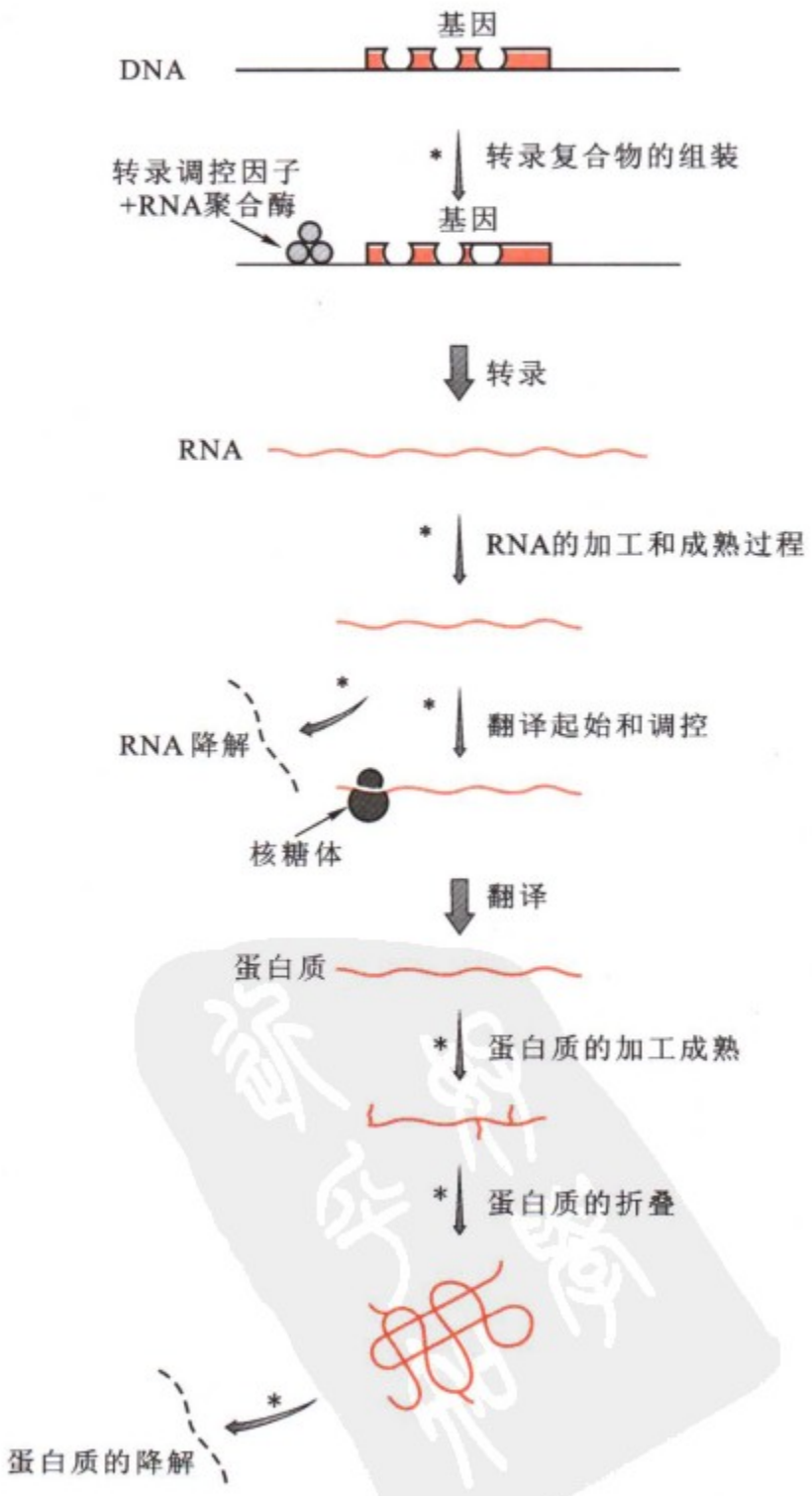



图 8-1 真核基因表达调控的主要步骤 * 调控位点

① 什么是诱发基因转录的信号?

② 基因调控主要是在哪一步(模板 DNA 的转录、mRNA 的成熟或蛋白质合成)实现的?

③ 不同水平基因调控的分子机制是什么?

## 8.1 真核生物的基因结构与转录活性

真核细胞与原核细胞在基因转录、翻译及 DNA 的空间结构方面存在如下几个方面的差异:

① 在真核细胞中,一条成熟的 mRNA 链只能翻译出一条多肽链,类似原核生物中常见的多基因操纵子形式不多。

② 真核细胞 DNA 与组蛋白和大量非组蛋白结合,只有一小部分 DNA 是裸露的。

③ 高等真核细胞 DNA 中很大一部分是不转录的,一些由几个或几十个碱基组成的 DNA 序列,在整个基因组中重复几百次甚至上百万次。此外,大部分真核细胞的基因中间还存在不被翻译的内含子。

④ 真核生物能够有序地根据生长发育阶段的需要进行 DNA 片段重排,还能在需要时增加细胞内某些基因的拷贝数,这种能力在原核生物中也是极为罕见的。

⑤ 在原核生物中,转录的调节区都很小,大都位于转录起始位点上游不远处,调控蛋白结合到调节位点上可直接促进或抑制 RNA 聚合酶对它的结合。在真核生物中,基因转录的调节区则大得多,它们可能远离核心启动子达几百个甚至上千个碱基对。虽然这些调节区也能与蛋白质结合,但是并不直接影响启动子区对于 RNA 聚合酶的接受程度,而是通过改变整个所控制基因 5′上游区 DNA 构型来影响它与 RNA 聚合酶的结合能力。

⑥ 真核生物的 RNA 在细胞核中合成,只有经转运穿过核膜,到达细胞质后,才能被翻译成蛋白质。原核生物中不存在这样严格的空间间隔。

⑦ 许多真核生物的基因只有经过复杂的成熟和剪接过程,才能被顺利地翻译成蛋白质。

### 8.1.1 基因家族

在原核细胞中,密切相关的基因往往组成操纵子,并且以多顺反子 mRNA 的方式进行转录,整个体系被置于一个启动子的控制之下。真核细胞的 DNA 是单顺反子结构,很少置于一个启动子控制之下的操纵子。真核细胞中许多相关的基因常按功能成套组合,被称为基因家族(gene family)。同一家族中的成员有时紧密地排列在一起,成为一个基因簇;更多的时候,它们却分散在同一染色体的不同部位,甚至位于不同的染色体上,具有各自不同的表达调控模式。

1. 简单多基因家族

简单多基因家族中的基因一般以串联方式前后相连。在大肠杆菌中,16 S,23 S 和 5 S *rRNA* 基因联合成一个转录单位,各种 rRNA 分子都是从这个转录

单位上剪切下来的(图 8-2)。事实上,细菌中所有 rRNA 和部分 tRNA 都来自这个沉降系数为 30 S(约 6 500 个核苷酸)的前 rRNA。原始转录产物首先被特异性甲基化,然后分别经 RNaseⅢ,RNaseP 和 RNaseE 在特定位点切开,生成 17 S、tRNA、35 S 和 5 S rRNA 前体分子,最后由特定核酸酶降解部分非必需序列,得到成熟的 16 S tRNA,23 S 和 5 S rRNA。参与 pre rRNA 降解成熟过程的 RNaseP 是一个核酶。基因组分析表明,大肠杆菌中共有 7 个这样的拷贝,每一个拷贝中只有 *tRNA* 基因的种类、数量和存在部位有些变化。个别拷贝中 5 S *rRNA* 基因下游还发现了部分共同转录的 *tRNA* 基因。

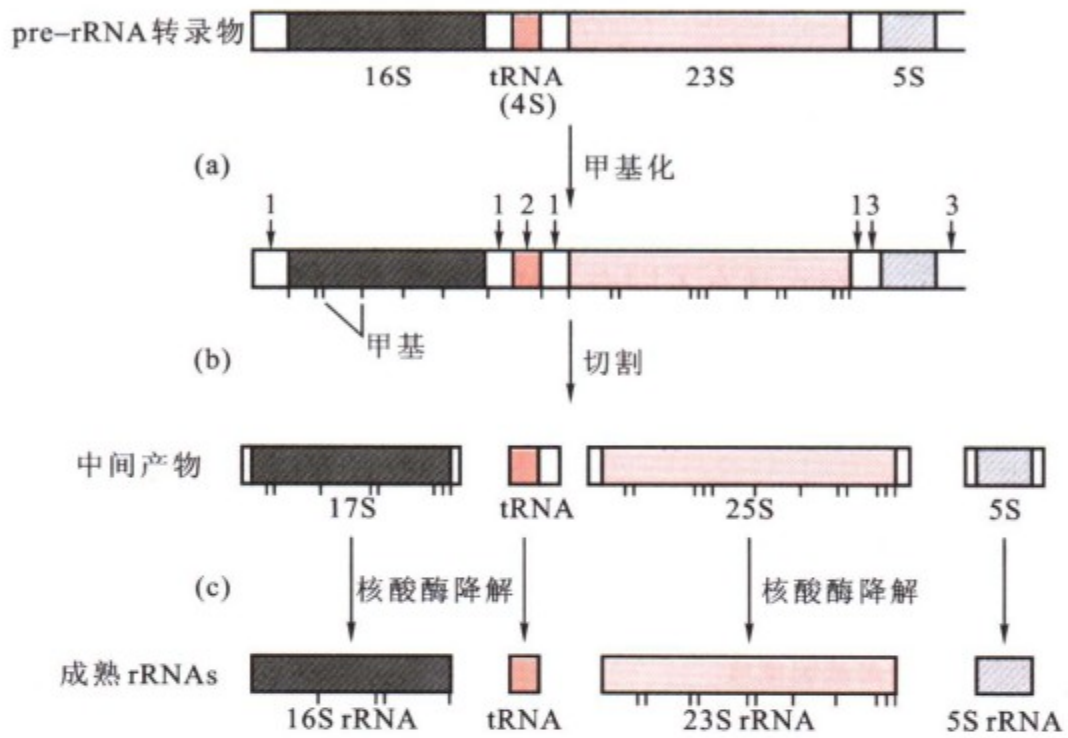



图 8-2 细菌中 *rRNA* 基因家族各成员的分布与成熟过程分析

在真核生物中,pre rRNA 转录产物的沉降系数为 45 S,(约有 14 000 个核苷酸),包括 18 S,28 S 和 5.8 S 三个主要 rRNA 分子。pre rRNA 分子中至少有 100 处被甲基化(主要是核糖的 2-OH 甲基化),原始转录产物也被特异性 RNA 酶切割降解,产生成熟 rRNA 分子。这个过程需要 snoRNAs 的参与(图 8-3)。5 S

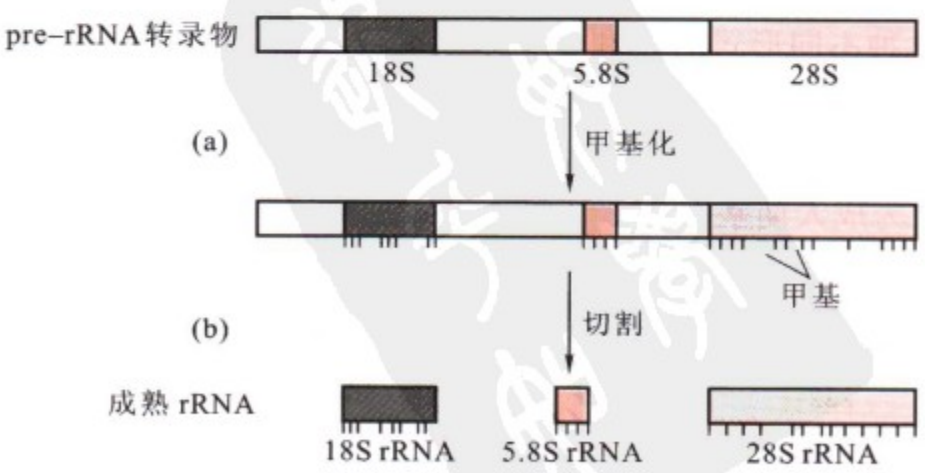



图 8-3 脊椎动物中 *rRNA* 基因家族主要成员的分布与成熟过程分析

rRNA 作为一个独立的转录单位,由 RNA 聚合酶Ⅲ(而不是聚合酶Ⅰ)完成转录。

2. 复杂多基因家族

复杂多基因家族一般由几个相关基因家族构成,基因家族之间由间隔序列隔开,并作为独立的转录单位。现已发现存在不同形式的复杂多基因家族。

我们先来看一下海胆的组蛋白基因家族(图 8-4)。在这种动物中,5 个分别编码不同组蛋白的基因处于一个约为 6 000 bp 的片段中,分别被间隔序列所隔开。这 5 个基因组成的串联单位在整个海胆基因组中可能重复多达 1 000 次。串联单位中的每一个基因分别被转录成单顺反子 RNA,这些 RNA 都没有内含子,而且各基因在同一条 DNA 链上按同一方向转录,每个基因的转录与翻译速度都受到调节。因为组蛋白只有在适合于染色体复制的情况下才大量合成,而且所合成的 $H_2A$、$H_2B$、$H_3$ 和 $H_4$ 的物质的量相等,$H_1$ 的量恰好是前者的一半。这是染色体组蛋白的实际比例。研究还表明,在一个特定的细胞中,并不是所有串联的单位都得到转录。在胚胎发育的不同阶段或不同组织中,有不同的串联单位被转录,暗示可能存在具有不同专一性的组蛋白亚类和发育调控机制。

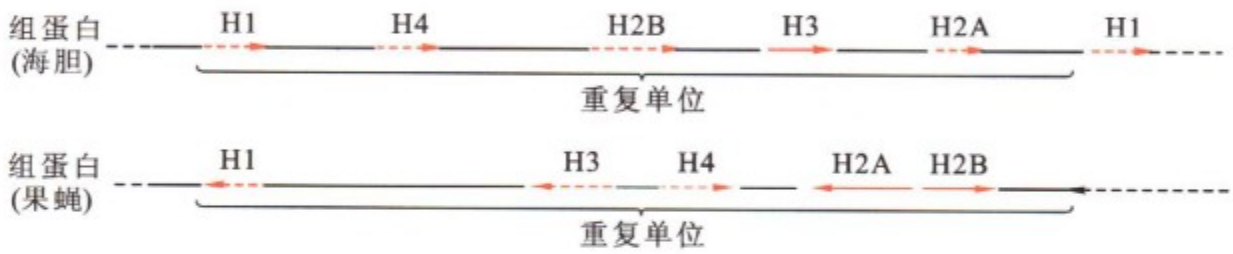



图 8-4 复杂的多基因家族

果蝇 *tRNA* 基因家族具有两个海胆组蛋白家族所没有的特点,那就是在基因家族中含有 3 个 $tRNA^{Lys}$ 基因,而且基因家族中的每个基因都单独按各自的方向进行转录。

3. 发育调控的复杂多基因家族

血红蛋白是所有动物体内输送分子氧的主要载体,由 2α2β 组成的四聚体加上一个血红素辅基(结合铁原子)后形成功能性血红蛋白。已知所有动物物种中血红蛋白基因的基本结构都相同(图 8-5)。然而,在生物个体发育的不同阶段,却出现几种不同形式的 α 和 β 亚基。人 α-珠蛋白基因簇位于 16 号染色体短臂上,约占 30 kb 左右,其中 ζ 为胚胎期基因(表 8-1)。β-珠蛋白基因簇位于 11 号染色体短臂上,约占 50~60 kb,其中 ε 为胚胎期基因,Gγ 和 Aγ 为胎儿型基因,δ 和 β 为成人期基因(图 8-6)。

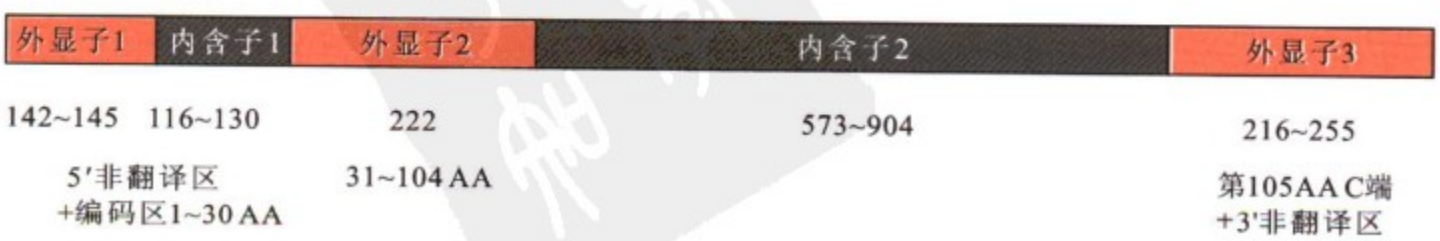



图 8-5 珠蛋白基因的基本结构(以人的 β-珠蛋白基因结构为例)

表 8-1 不同发育阶段血红蛋白亚型

| 发育阶段 | 组成 |
| --- | --- |
| 胚胎期(8 周以前) | $\zeta_2\varepsilon_2$、$\zeta_2\gamma_2$ 和 $\alpha_2\varepsilon_2$ |
| 胎儿期(8 ~ 41 周) | $\alpha_2\gamma_2$ |
| 成人期(出生以前) | $\alpha_2\delta_2$ 和 $\alpha_2\beta_2$ |

胚胎期类 α-珠蛋白，最先以 $\zeta_2$ 型出现，以后逐渐被 $\zeta_1$ 型所取代。在胎儿期有两类等分子数 β 型链，即 Gγ 型和 Aγ 型链(在肽链的同一位点上，一条链含有甘氨酸，另一条含有丙氨酸)。因此，胚胎型血红蛋白中含有两个 $\zeta_2$ 型 α 链和两个 $\varepsilon_2$ 型 β 链(即 $2\zeta_2 2\varepsilon_2$)。新生儿出生后，98%的血红蛋白中含有两个 $\alpha_2$ 亚单位和两个 β 亚单位，2%的血红蛋白含两个 $\alpha_2$ 亚单位和两个 δ 亚单位。在每个基因家族中，基因排列的顺序就是它们在发育阶段的表达顺序(图 8-7)。

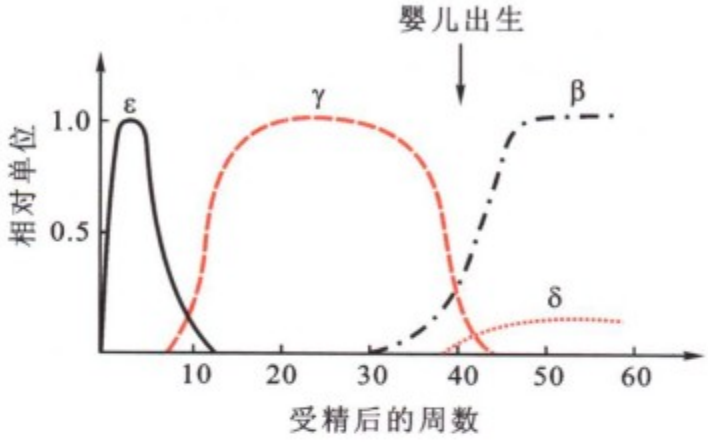



图 8-6 人体发育过程中不同类型 β-珠蛋白的含量变化

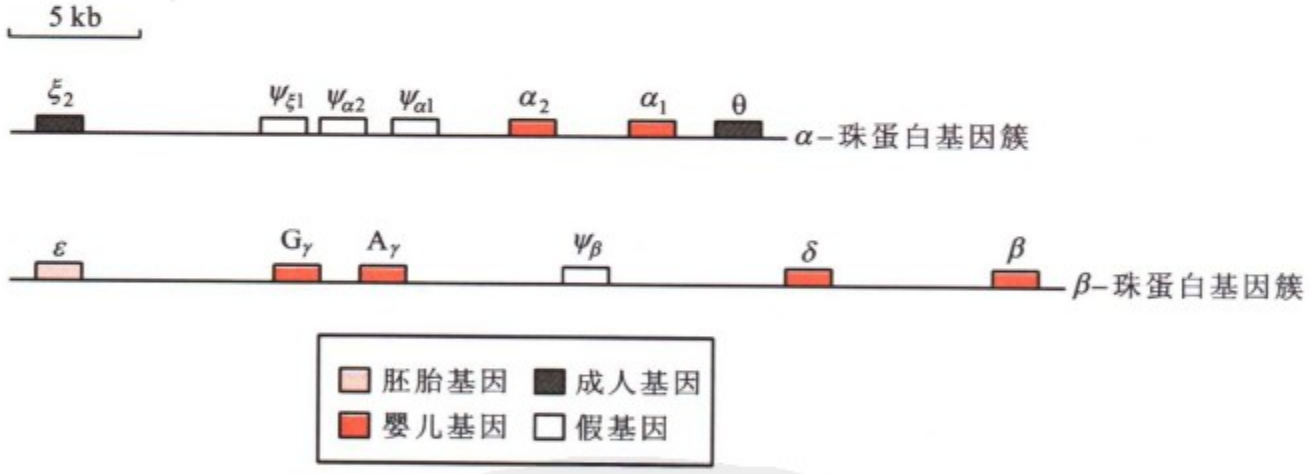



图 8-7 人细胞中 α 和 β-珠蛋白基因簇结构

原始鱼类、海蚕及昆虫只有单个珠蛋白基因，两栖类的 α 和 β 基因则紧密地连锁在同一条染色体上，而低等哺乳动物和鸟类的 α 和 β 基因有多种亚型，基因家族分别位于不同的染色体上。科学家认为，原始珠蛋白大约产生于 8 亿年前，由单基因编码，现代珠蛋白基因家族是由同一个原始基因通过一系列重复、突变和转位演变而成的。原始生物的体积都很小，虽然只有一个血红蛋白亚基，运氧能力有限，但也能满足生理功能的需要。随着物种的进化，动物的体积相对变大，对氧的需求量逐渐增加，在这一进化压力下，血红蛋白基因通过自发突变产生杂合基因，编码出对氧的结合和释放能力要大大高于单分子血红蛋白的四聚体。在哺乳动物的进化过程中，两条 β 链基因又发生突变和重复，形成了

胎儿中 ε 和 γ 型珠蛋白。胎儿血红蛋白对氧的亲和力比成人更大,因而有利于胎儿的快速发育。

## 8.1.2 真核基因的断裂结构

1. 外显子与内含子

Gilbert 第一个用"intron"来描述存在于原始转录产物或基因组 DNA 中,但不包括在成熟 mRNA、rRNA 或 tRNA 中的那部分核苷酸序列。现已查明,大多数真核基因都是由蛋白质编码序列和非蛋白质编码序列两部分组成的(表 8-2)。编码序列称为外显子(exon),非编码序列称为内含子(intron)。在一个结构基因中,编码某一蛋白质不同区域的各个外显子并不连续排列在一起,而是常常被长度不等的内含子所隔离,形成镶嵌排列的断裂方式。所以,真核基因有时被称为断裂基因(interrupted gene)。不同基因拥有数量和大小都不同的内含子,如胶原蛋白基因长约 40 kb,至少有 40 个内含子,其中短的只有 50 bp,长的可达到 2 000 bp。

表 8-2 不同真核生物基因的平均长度及单个基因平均含有外显子数量比较

| 物种 | 外显子数/基因 | 核基因平均长度(kb) | mRNA 平均长度 |
|---|---|---|---|
| 酵母 | 1 | 1.6 | 1.6 |
| 真菌 | 3 | 1.5 | 1.5 |
| 线虫 | 4 | 4.0 | 3.0 |
| 果蝇 | 4 | 11.3 | 2.7 |
| 鸡 | 9 | 13.9 | 2.4 |
| 哺乳动物 | 7 | 16.6 | 2.2 |

许多情况下,高等真核基因核 DNA 序列中对应于外显子的部分可能还不到 10%。图 8-8 是哺乳动物二氢叶酸还原酶基因,全长 25~31 kb,但其 6 个外显子总长只有 2 kb。少数基因,如组蛋白及 α 型、β 型干扰素基因,根本不包括内含子。目前尚不清楚内含子的生理功能。在某些基因中,完全除去内含子以后,照样可以产生有活性的 mRNA;而另一些基因,如 SV40 T 抗原基因,一旦除去

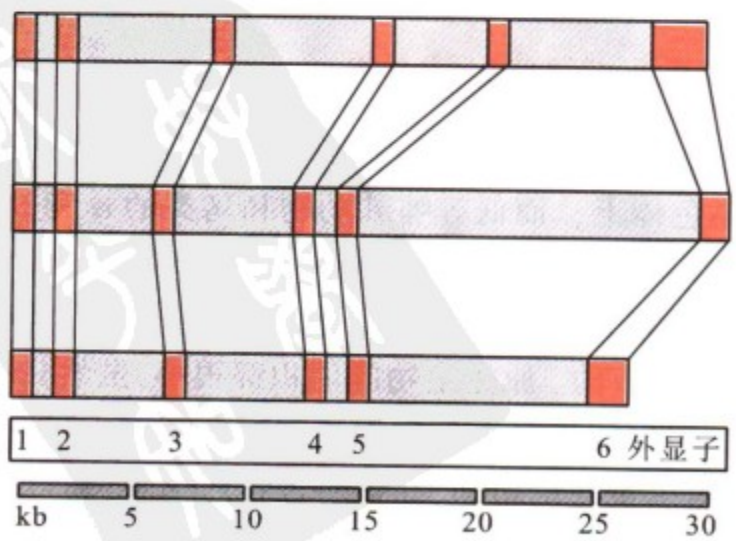



图 8-8 哺乳动物二氢叶酸还原酶的基因结构

内含子,成熟 mRNA 运入细胞质的过程就完全被阻断。研究发现,只有真核生物具有切除基因中内含子,产生功能型 mRNA 和蛋白质的能力,原核生物一般不具有这种本领。如果要在原核细胞里表达真核基因,必须首先构建切除内含子的重组基因,才有可能得到所研究的蛋白质。

断裂基因的结构形式为某些重要蛋白质的编码区域提供了进行重组的潜在位点,使这些 DNA 序列有充分的机会进行重复和组合,非常有利于真核基因的进化。

2. 外显子与内含子的连接区

真核基因断裂结构的另一个重要特点是外显子-内含子连接区(exon-intron junction)的高度保守性和特异性碱基序列。外显子-内含子连接区就是指外显子和内含子的交界或称边界序列,它有两个重要特征:内含子的两端序列之间没有广泛的同源性,因此内含子两端序列不能互补,说明在剪接加工之前,内含子上游序列和下游序列不可能通过碱基配对形成发卡式二级结构。外显子-内含子连接区序列虽然很短,但却是高度保守的,因此,该序列可能与剪接机制密切相关,是 RNA 剪接的信号序列。我们将在转录初级产物的加工这一节中做详细介绍。

序列分析表明,几乎每个内含子 5′端起始的两个碱基都是 GT,而 3′端最后两个碱基总是 AG,由于这两个碱基的高度保守性和广泛性,有人把它称为 GT-AG 法则,即:5′GT……AG 3′。由于内含子两末端的序列不同,可定向标明内含子的两个末端,根据剪接加工过程沿内含子自左向右进行的原则,一般将内含子 5′端接头序列称为左剪接位点,3′端接头序列称为右剪接位点,有时也将前者称为供体位点(donor site),将后者称为受体位点(acceptor site)。

外显子-内含子在连接区的保守序列几乎存在于所有高等真核生物基因中,表明可能存在着共同的剪接加工机制。但是在线粒体基因和酵母 *tRNA* 基因中不存在这类保守序列,暗示存在着不同类型的加工剪接过程。

3. 外显子与内含子的可变调控

在基因转录,加工产生成熟 mRNA 分子时,内含子通过剪接加工被去掉,保留在成熟 mRNA 分子中的外显子被拼接在一起,最终被翻译成蛋白质。因此通过反转录酶的作用,由成熟 mRNA 产生的 cDNA 分子中,只含有外显子,没有内含子。

原始转录产物内含子与外显子边界序列在剪接酶或 RNA 本身的作用下,发生的磷酸二酯键断裂,并产生连接两个外显子的化学键。有些真核基因,如肌红蛋白重链基因虽有 41 个外显子,却能精确地剪接成一个成熟的 mRNA,我们称这种方式为组成型剪接。一个基因的转录产物通过组成型剪接只能产生一种成熟的 mRNA,编码一个多肽。但是,有不少真核基因的原始转录产物可通过不同的剪接方式,产生不同的 mRNA,并翻译成不同的蛋白质。另外一些核基因由于转录时选择了不同的启动子,或者在转录产物上选择不同的 poly A 位点而使原始转录物具有不同的二级结构,因而影响剪接过程,最终产生不同的 mRNA 分子。同一基因的转录产物由于不同的剪接方式形成不同 mRNA 的过程称为选择性剪接。

小鼠淀粉酶基因表达的组织特异性变化就是一个例子(图 8-9)。已知在肝脏和唾液腺中都合成这个酶,由同一个基因编码。在这两种组织中,淀粉酶

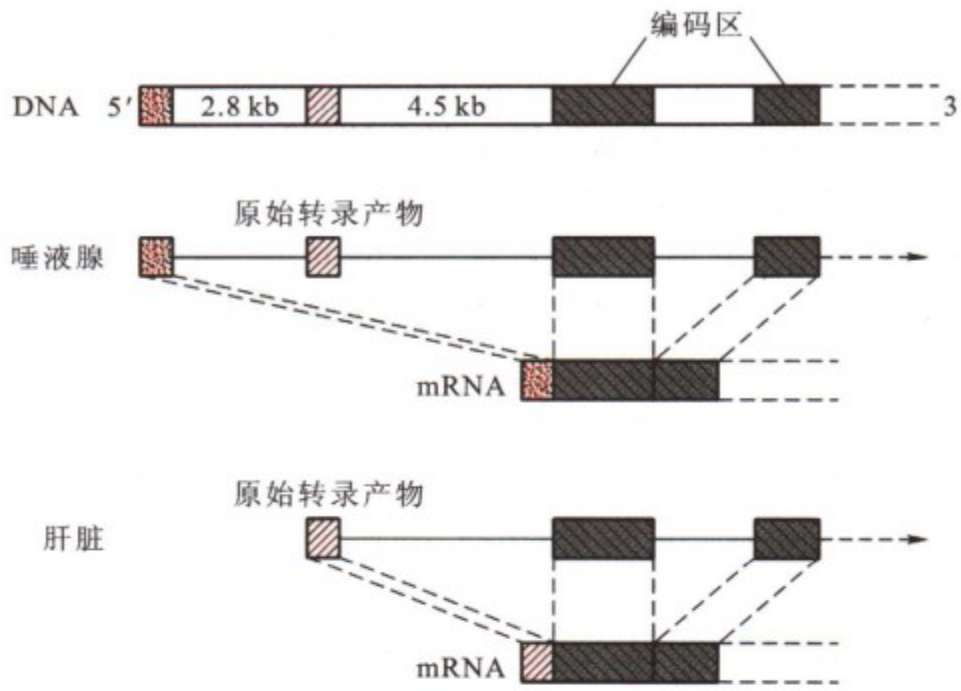



图 8-9 不同外显子的使用导致 α-淀粉酶基因在不同组织中的表达差异

mRNA 的编码序列完全相同,只是 5′端起始部分长度不同。它的编码序列起始于第 2 号外显子,该外显子与其余一系列外显子相连,形成完整的编码序列。不同组织中不编码蛋白质的第一号外显子的序列却不相同。在肝脏中,mRNA 5′端的 161 个碱基是由位于第 2 号外显子转录起始点上游 4 500 bp 处的 L 外显子编码的。在唾液腺中,mRNA 5′端的 50 个碱基是由位于转录起始点上游 7 300 bp 处的 S 外显子编码的。这样,L 外显子和 S 外显子分别为淀粉酶 mRNA 提供了不同的起始序列。事实上,L 外显子只是唾液腺淀粉酶基因中内含子序列的一部分,将在mRNA 成熟过程中被切除。有实验证明,由 S 外显子起始的转录产物是由 L 外显子起始转录产物的 100 倍以上。由于一个基因的内含子成为另一个基因的外显子,形成基因的差别表达,这是真核基因断裂结构的一个重要特点。

### 8.1.3 真核生物 DNA 水平上的基因表达调控

分子生物学的最新研究表明,在个体发育过程中,用来合成 RNA 的 DNA 模板也会发生规律性变化,从而控制基因表达和生物体的发育。

高度重复基因的形成通常与个体分化阶段 DNA 的某些变化有关。例如,一个成熟的红细胞能产生大量的可翻译出成熟珠蛋白的 mRNA,而其前体细胞却不产生珠蛋白。许多情况下,这种变化是由于基因本身或它的拷贝数发生了永久性变化。这种 DNA 水平的调控是真核生物发育调控的一种形式,它包括了基因丢失、扩增、重排和移位等方式,通过这些方式可以消除或变换某些基因并改变它们的活性。这些调控方式与转录及翻译水平的调控是不同的,因为它使基因组发生了改变。

1. “开放”型活性染色质(active chromatin)结构对转录的影响

真核基因的活跃转录是在常染色质上进行的。转录发生之前,染色质常常会在特定的区域被解旋松弛,形成自由 DNA。这种变化可能包括核小体结构的

消除或改变,DNA 本身局部结构的变化,从右旋型变为左旋型(Z-DNA)等,这些变化可导致结构基因暴露,促进转录因子与启动区 DNA 的结合,诱发基因转录。

用 DNA 酶 I 处理各种组织的染色质时,发现处于活跃状态的基因比非活跃状态的 DNA 更容易被 DNA 酶 I 所降解。鸡成红细胞(erythroblast)染色质中,β-血红蛋白基因比卵清蛋白基因更容易被 DNA 酶 I 切割降解。与此相反,鸡输卵管细胞的染色质中被 DNA 酶 I 优先降解的是卵清蛋白基因,而不是β-血红蛋白基因。那么,活跃状态的 DNA 为什么更易于受核酸酶的攻击而降解呢?研究发现,活跃表达基因所在染色质上一般含有一个或数个 DNA 酶 I 超敏感位点(hypersensitive site),它们大多位于基因 5′端启动区,少数在其他位置。非活性态基因的 5′端相应位点却不表现对 DNA 酶 I 的超敏感性。果蝇唾液腺染色体中,胶原蛋白基因 *sgs4* 的转录区上游 330 bp 和 405 bp 左右各有一个超敏感位点,有突变体缺失了包括这两个超敏感位点大约 100 个 bp 的 DNA 序列后就不再具备合成胶原蛋白的能力,说明两个完整的超敏感位点是 *sgs4* 基因转录所必需的。

此外,鸡成红细胞中β-血红蛋白基因 5′端-280~-50 bp 区是超敏感位点。有人用专一切割单链 DNA 的 SI 核酸酶处理该基因活跃表达的染色体 DNA,证实有 DNA 被水解,说明该基因活跃表达时启动区部分序列可能解开成单链,从而不能继续缠绕在核小体上,使启动区 DNA“裸露”于组蛋白表面,形成了对 DNA 酶 I 的超敏感现象。上述事实充分说明,超敏感位点的产生可能是染色质结构规律性变化的结果。正是由于这种变化,使 DNA 容易与 RNA 聚合酶和其他转录调控因子相结合,从而启动基因表达,同时也更易于被核酸酶所降解。

有证据表明,存在于“灯刷型”染色体(lamp brush)上的环形结构可能与基因的活性转录有关(图 8-10)。“灯刷型”染色体只有在两栖动物卵细胞发生减数分裂时才能被观察到,它是染色体充分伸展时的一种形态。此时,两对姐妹染色体常常通过“交叉点”(chiasmata)连成一体(图 8-10a)。高倍电镜下观察发现,灯刷型染色体上存在许多突起的“泡”状或“环”状结构,有时还能看到 RNP 沿

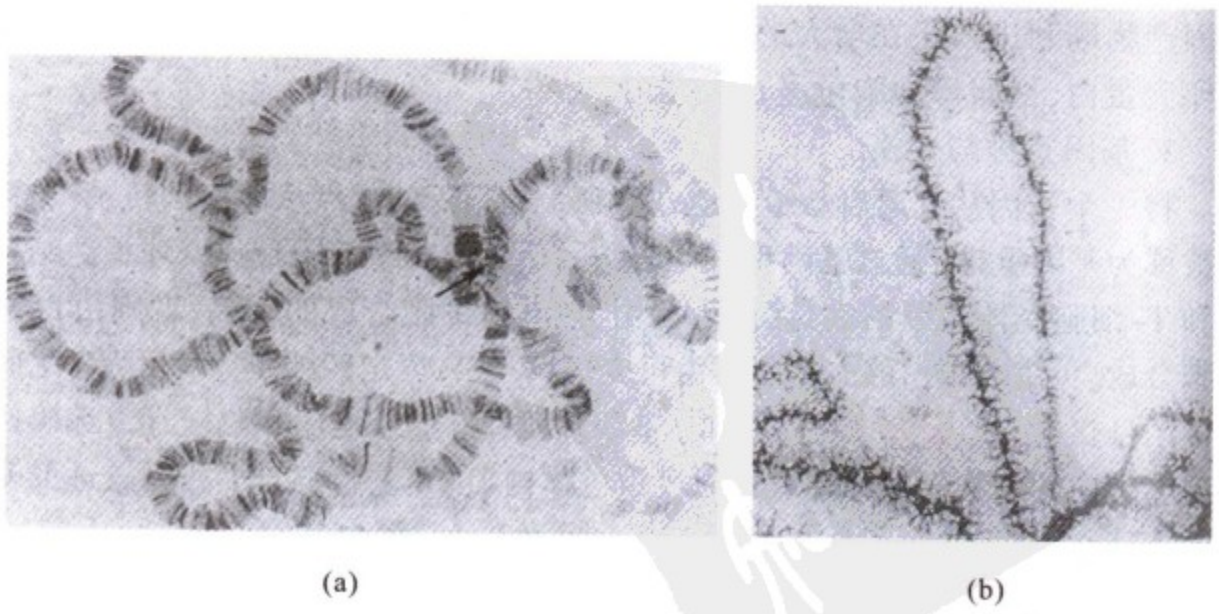



图 8-10 “灯刷型”染色体(a)及存在于“活性”染色质区的环状 DNA 结构(b)(a)中箭头所示为不同姐妹染色体对之间的交叉点

着这些突起结构移动,表明这些 DNA 正在被 RNA 聚合酶所转录(图 8-10b)。$^3H$ 尿嘧啶核苷酸标记实验表明,最活跃的 DNA 转录区结构较为伸展,形成蓬松区。

研究小鼠细胞非活化染色质时发现,核小体核心组蛋白 H3 第 110 位的巯基处于封闭状态。当该基因活跃表达时,核小体转变为伸展状态,H3 上可检测到巯基活性。研究原癌基因 *c-fos* 和 *c-myc* 的诱导表达时也发现,在基因活跃转录的数十分钟或数小时内,这些基因绝大部分都分布在伸展状态(即活性)染色质中,而一旦转录终止,这些基因又重新分布到非活化的压缩状态中,表明转录时核小体伸展或压缩状态的转换是可逆的。

2. 基因扩增

基因扩增是指某些基因的拷贝数专一性大量增加的现象,它使细胞在短期内产生大量的基因产物以满足生长发育的需要,是基因活性调控的一种方式。例如,非洲爪蟾的卵母细胞中原有 *rRNA* 基因(rDNA)约 500 拷贝,在减数分裂Ⅰ的粗线期,这个基因开始迅速复制,到双线期它的拷贝数约为 200 万,扩增近4 000 倍,可用于合成 $10^{12}$ 个核糖体,以满足卵裂期和胚胎期合成大量蛋白质的需要。在基因扩增之前,整个 rDNA 区位于单个核仁中(人们发现大部分动物细胞中都有这种含 rDNA 的核内小体)。在该细胞发育成卵母细胞的过程中(约 3 周时间),不但 rDNA 数量猛增,核仁数量也大大增加,每个核中可有几百个这样的小核仁。一旦卵母细胞成熟,多余的 rDNA 就没有用了,将被逐渐降解。受精之后,染色体 DNA 开始复制,并通过有丝分裂的方式,不断扩大细胞群体。在此期间,多余的 rDNA 继续被降解,直到分裂产生几百个细胞时,rDNA 的过剩现象就不复存在了。

果蝇中也发现了基因扩增现象。在卵巢成熟之前,卵巢颗粒细胞中产生卵壳蛋白的基因被扩增。果蝇中的卵原细胞,经 4 次分裂产生 16 个细胞,其中一个是卵母细胞(oocyte),将发育成为卵细胞。其他 15 个是营养细胞(nurse cell),它们为卵细胞的形成提供大量的蛋白质及其他大分子物质。营养细胞之所以能够产生大量的营养物质,是因为在它们的形成过程中发生了多次特殊的 DNA 复制,卵壳蛋白等基因拷贝数显著增加。营养细胞与卵细胞通过胞质桥相连,使基因产物顺利进入正在增大的卵细胞,因此,在卵发生后期,营养细胞能合成大量卵壳蛋白,把卵母细胞包裹在卵壳内。

3. 基因重排与变换

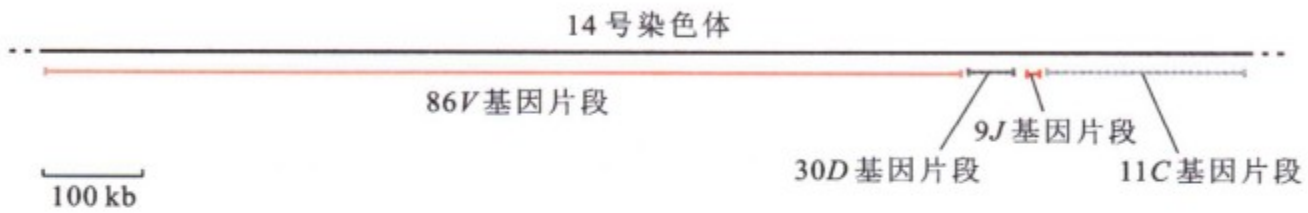
将一个基因从远离启动子的地方移到距它很近的位点从而启动转录,这种方式被称为基因重排。通过基因重排调节基因活性的典型例子是免疫球蛋白结构基因和 T-细胞受体基因的表达,前者是由 B-淋巴细胞合成的,而后者则由T-淋巴细胞合成。免疫球蛋白的肽链主要由可变区(V 区)、恒定区(C 区)以及两者之间的连接区(J 区)组成(表 8-3),*V*、*C* 和 *J* 基因片段在胚胎细胞中相隔较远。编码产生免疫球蛋白的细胞发育分化时,通过染色体内 DNA 重组把 4 个相隔较远的基因片段连接在一起,从而产生了具有表达活性的免疫球蛋白基因(图 8-11)。

酵母有两种交配型,*MATa* 和 *MAT*α,相当于高等真核生物中的雌性和雄性。由于酵母主要通过无性的细胞分裂方式来增殖,整个群体很容易变成只有一种交配型,从而失去通过有性繁殖得到下一代的机会。这种现象在遗传学上

表 8-3 人类基因组中免疫球蛋白基因主要片段的数量比较

| 成份 | 基因位点 | 染色体 | 基因片段数量 | | | |
|---|---|---|---|---|---|---|
| | | | V | D | J | C |
| 重链 | IGH | 14 | 86 | 30 | 9 | 11 |
| 轻链(κ 链) | IGK | 2 | 76 | 0 | 5 | 1 |
| 轻链(λ 链) | IGL | 22 | 52 | 0 | 7 | 7 |

(a) IGH 基因座的分布情况




(b) 通过基因重排 (rearrangement) 获得功能型免疫球蛋白重链基因

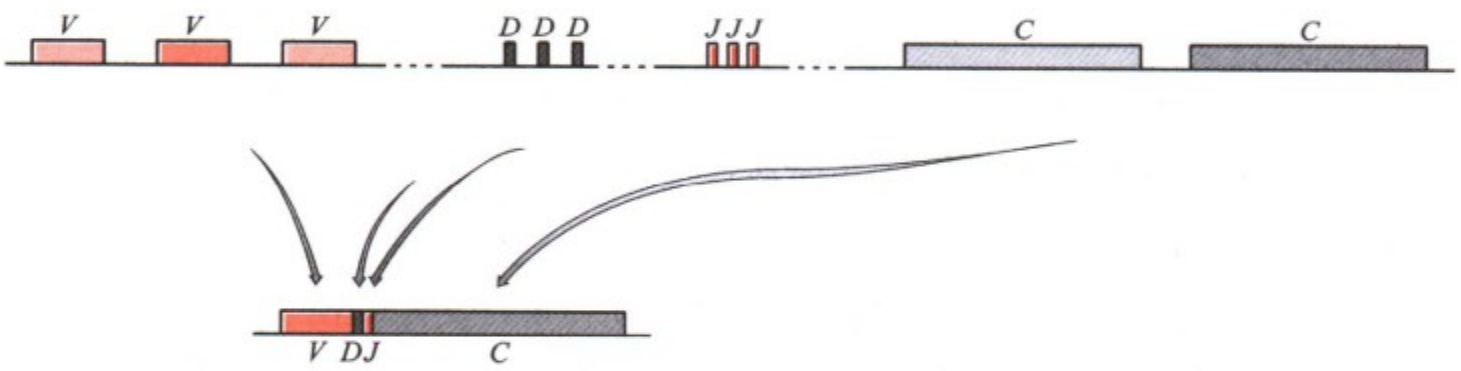



图 8-11 免疫球蛋白重链基因片段重排与组织特异性表达

是有风险的,因为无性生殖得到的群体中遗传性状往往较为单一。为了避免这种情况,酵母细胞能通过一种被称为“交配型转换”(mating type switching)的过程改变自己的性别。这个过程有时也被称为基因转换(gene conversion)。研究酵母染色体上交配型基因位点发现,每个单倍体酵母细胞都分别带有 *MATa* 或 *MATα* 基因,在这个基因位点的两侧分别存在着两个 *MAT* 样基因,*HMLα* 和 *HMRa*,其序列分别与 *MATα* 和 *MATa* 完全相同,只不过这两个基因一般情况下都不能表达,在它们的上游调控区有“沉默子”阻遏基因表达。所以,*HMLα* 和 *HMRa* 又被称为 “沉默交配型盒”。交配型转换的第一步是核酸内切酶 HO 在 *MAT* 基因内部一个 24 bp 序列处切开双链 DNA,自由的 3′端又在 DNA 聚合酶的作用下以上游(或下游)*HMLα* 基因为模板合成双链 DNA,直到完全置换原有的 *MATa* 基因,完成性别变换(图 8-12)。

此外,人血红蛋白基因表达也有这种现象。β 链基因家族的调节基因和受体基因原来是被间隔序列分隔开的,因此这些结构基因没有活性。在胚胎发育早期,由于间隔序列 $S_1$ 缺失,导致调节基因 $R_1$ 和受体基因 $A_1$ 连接在一起,形成调节启动基因($R_1A_1$),邻近的结构基因转录产生 ε 肽链。在胚胎发育后期(3~9

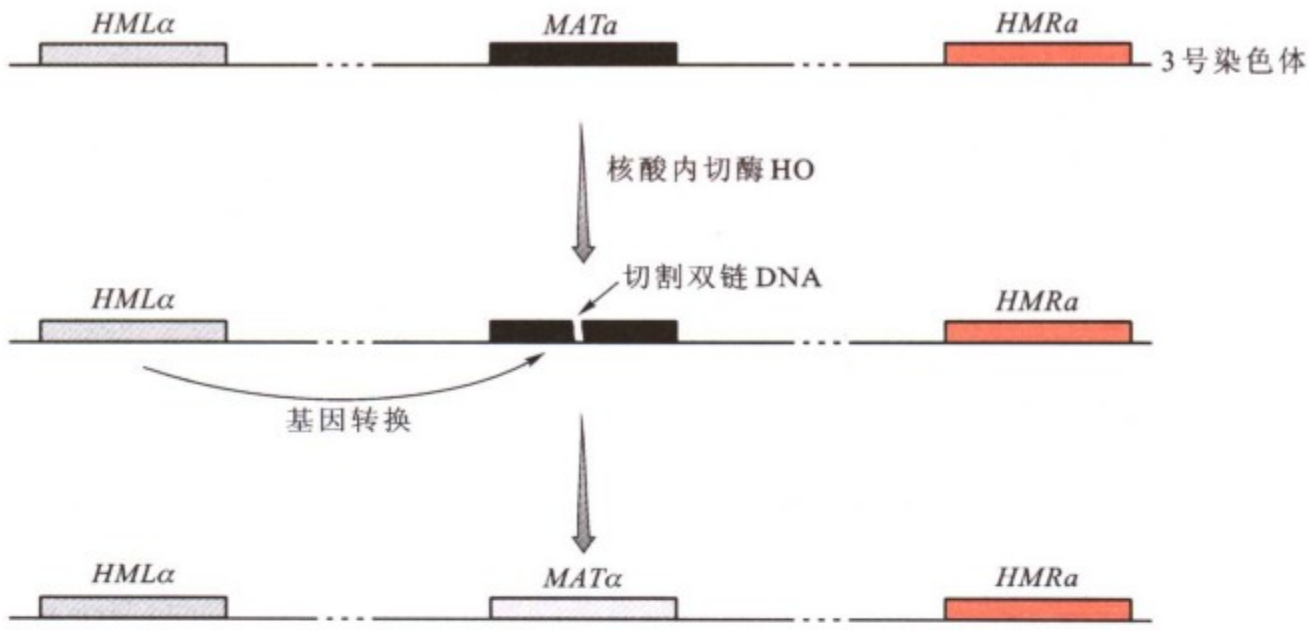



图 8-12 酵母基因重排与交配型转换

个月),由于间隔序列 $S_2$ 缺失,使调节基因 $R_2$ 和 $A_2$ 连接在一起,形成调节启动基因($R_2A_2$),邻近的结构基因转录产生 γ 肽链。人出生以后,由于 $S_3$ 的缺失,形成了 $R_3A_3$ 调节启动基因,激活 β 和 δ 结构基因转录,产生 β 和 δ 肽链。α 肽链基因的变换也按这个模式进行,形成了人类血红蛋白肽链组成的规律性变化。

### 8.1.4 DNA 甲基化与基因活性的调控

DNA 甲基化是最早发现的修饰途径之一,这一修饰途径可能存在于所有高等生物中并与基因表达调控密切相关。大量研究表明,DNA 甲基化能关闭某些基因的活性,去甲基化则诱导了基因的重新活化和表达。DNA 甲基化能引起染色质结构、DNA 构象、DNA 稳定性及 DNA 与蛋白质相互作用方式的改变,从而控制基因表达。研究证实,CpG 二核苷酸中胞嘧啶的甲基化导致了人体 1/3 以上由于碱基转换而引起的遗传病。我们将在本节中对 DNA 甲基化和去甲基化的机制、DNA 甲基化对发育过程中基因表达的调控以及与之有关的研究进展一一简述。

1. DNA 的甲基化

DNA 甲基化修饰现象广泛存在于多种有机体中。实验证明,这个过程不但与 DNA 复制起始及错误修正时的定位有关,还通过改变基因的表达参与细胞的生长、发育过程及染色体印迹、X 染色体失活等的调控。DNA 甲基化主要形成 5-甲基胞嘧啶(5-mC)和少量的 $N^6$-甲基腺嘌呤($N^6$-mA)及 7-甲基鸟嘌呤(7-mG, 图 8-13)。在真核生物中,5-甲基胞嘧啶主要出现在CpG 序列、CpXpG、CCA/TGG 和 GATC 中。因为高等生物 CpG 二核苷酸序列中的 C 通常是甲基化的,极易自发脱氨,生成胸腺嘧啶,所以 CpG 二核苷酸序列出现的频率远远低于按核苷酸组成计算出的频率。由于这些 CpG 二核苷酸通常成串出现在 DNA 上,因此这段序列往往被称为 CpG 岛。

真核生物细胞内存在两种甲基化酶:一种被称为日常型甲基转移酶,另一种是从头合成型甲基转移酶,前者主要在甲基化母链(模板链)指导下使处于半

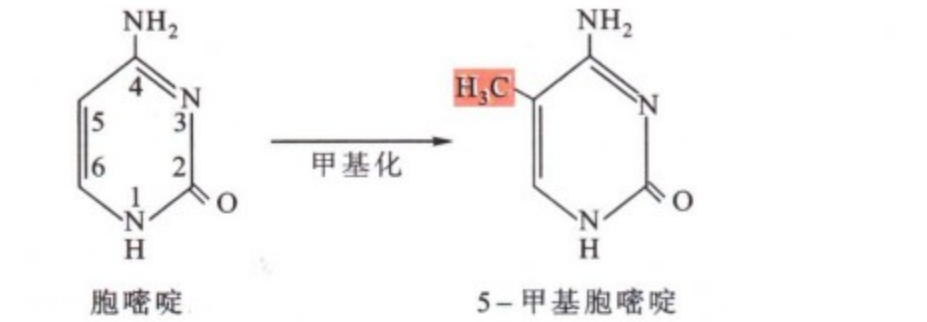



图 8-13 5-甲基胞嘧啶、$N^6$-甲基腺嘌呤和 7-甲基鸟嘌呤结构式

甲基化的 DNA 双链分子上与甲基胞嘧啶相对应的胞嘧啶甲基化。该酶催化特异性极强，对半甲基化的 DNA 有较高的亲和力，使新生的半甲基化 DNA 迅速甲基化，从而保证 DNA 复制及细胞分裂后甲基化模式不变。后者催化未甲基化的 CpG 成为 mCpG，它不需要母链指导，但速度很慢。

2. DNA 甲基化抑制基因转录的机制

DNA 甲基化导致某些区域 DNA 构象变化，从而影响了蛋白质与 DNA 的相互作用，抑制了转录因子与启动区 DNA 的结合效率。研究表明，当组蛋白 H1 与含 CCGG 序列的甲基化或非甲基化 DNA 分别形成复合体时，DNA 的构型存在着很大的差别，甲基化达到一定程度时会发生从常规的 B-DNA 向 Z-DNA 的过渡。由于 Z-DNA 结构收缩，螺旋加深，使许多蛋白质因子赖以结合的元件缩入大沟而不利于基因转录的起始。有实验用序列相同但甲基化水平不同的 DNA 为材料，比较其作为 RNA 聚合酶转录模板的活性，发现甲基的引入不利于模板与 RNA 聚合酶的结合，降低了其体外转录活性。

5-甲基胞嘧啶在 DNA 上并不是随机分布的，基因的 5′端和 3′端往往富含甲基化位点，而启动区 DNA 分子上的甲基化密度与基因转录受抑制的程度密切相关(图 8-14)。对于弱启动子来说，稀少的甲基化就能使其完全失去转录活性。当这一类启动子被增强时(带有增强子)，即使不去甲基化也可以恢复其转

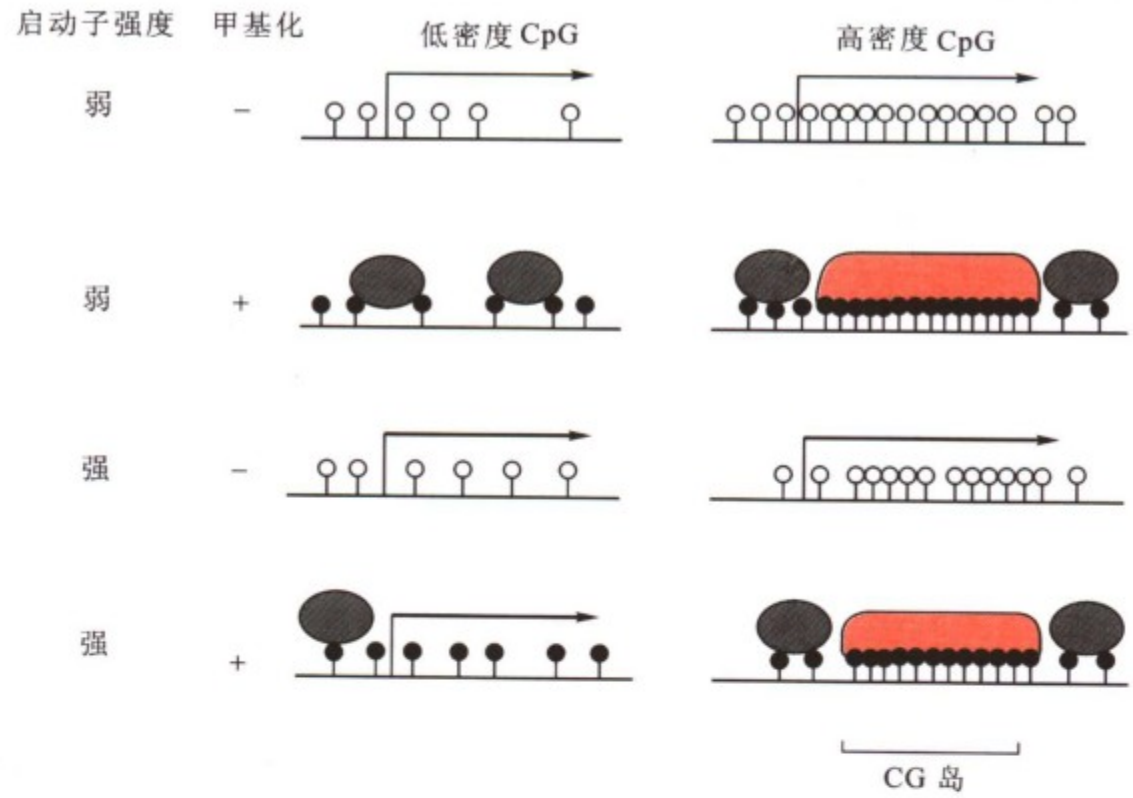



**图 8-14 DNA 甲基化对基因转录的抑制作用** 箭头粗细表示转录活性强弱,空心圆表示 DNA 未甲基化,小黑点表示 DNA甲基化,椭圆形表示 MeCPl 松散结合,长方形表示 MeCPl 紧密结合。

录活性。若进一步提高甲基化密度,即使增强后的启动子仍无转录活性。因为甲基化对转录的抑制强度与 MeCPl(methyl CpG-binding protein 1)结合 DNA 的能力呈正相关,甲基化 CpG 的密度和启动子强度之间的平衡决定了该启动子是否具有转录活性。

研究证实,小鼠 α-珠蛋白基因和人 γ-珠蛋白基因内 CpG 密度较低,前者 1.4kb 序列内含有 28 个 CpG(相当于每 50 bp 有一个 CpG),后者 3.3 kb 序列内含有 26 个 CpG(相当于每 126 bp 有一个 CpG),而人 α-珠蛋白基因内 CpG 密度很高,1.5 kb 序列中有 141 个 CpG (相当于每 10bp 有一个 CpG)(图 8-15a)。可以用细菌 *Hha* Ⅰ甲基转移酶处理人 α-珠蛋白基因使其低水平甲基化(平均每 62 个bp 有一个 mCpG,1.5 kb 序列中有 24 个 mCpG), 也可以用 *Sss* Ⅰ甲基化酶把该基因上所有 CpG 全部甲基化。将上述 CpG 密度不同、甲基化水平也不同的基因导入人 HeLa 细胞发现,一旦 CpG 完全甲基化,这 3 个基因都不表达。但如果在这些基因的启动区装上 SV40 增强子序列,CpG 密度较低的基因如小鼠 α-珠蛋白基因和人 γ-珠蛋白基因就得到表达, 高 CpG 密度但只有部分甲基化的人 α-珠蛋白基因也能被转录,CpG 密度高同时又完全甲基化的人 α-珠蛋白基因仍然不被转录(图 8-15b)。

DNA 的甲基化还提高了该位点的突变频率。真核生物中 5-mC 主要出现在 5′-CpG-3′序列中,5-mC 脱氨后生成的胸腺嘧啶(T),不易被识别和校正。因此,特定部位的 5-mC 脱氨基反应,将在 DNA 分子中引入可遗传的转化(C→T),若位点突变发生在 DNA 功能区域,就可能造成基因表达的紊乱。在脑瘤、乳腺癌和直肠癌细胞中,*p53* 基因第 273 位密码子含 CpG 序列,常由 CGT 突变为 CAT

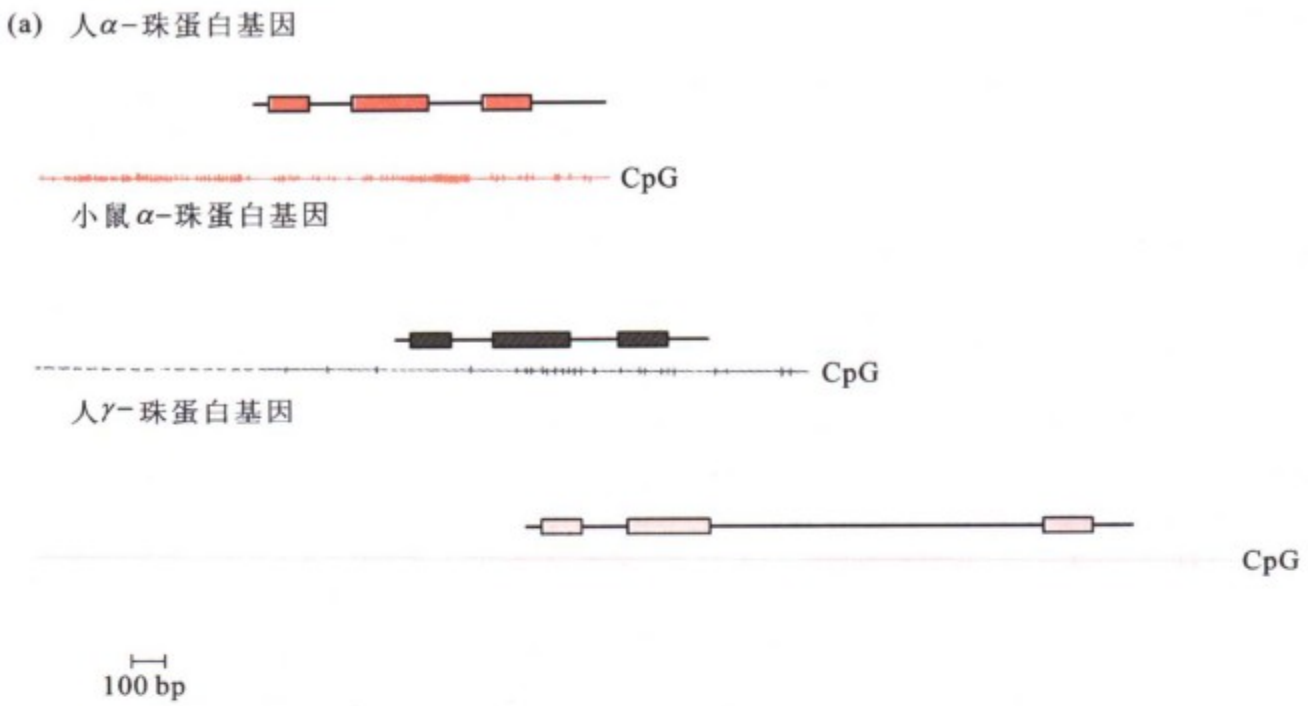



**图 8-15 三个珠蛋白的基因编码区及启动区的 CpG 密度比较(a)和甲基化对这些基因表达强高的影响(b)** 无:没有甲基化;部:部分甲基化;β-珠蛋白mRNA 作为对照。

或 TGT(Arg→His 或 Cys)。在非小细胞肺癌中 *p53* 基因该位点 G→T 的突变频率高达 59.3%。由于 CpG 甲基化增加了胞嘧啶残基突变的可能性,5-mC 也作为内源性诱变剂或致癌因子调节基因表达。

3. DNA 甲基化与 X 染色体失活

X 染色体失活是发育过程中独特的调节机制。雌性胎生哺乳动物细胞中两条 X 染色体之一在发育早期随机失活,以确保与只有一条 X 染色体的雄性个体内 X 染色体基因的剂量相同。一旦发生 X 染色体失活,这个信息便能够稳定地传递给子代细胞,使该细胞有丝分裂所产生的后代都保持同一条 X 染色体失活。因为人类 X 染色体上有 1 亿 2 千 800 万个碱基对和 1 465 个基因,所以 X 染色体失活机制与大量基因的表达与否密切相关,是发育调控中的一个广为研

究的课题。

很早以前，人们就发现X染色体上存在一个与X染色体失活有密切联系的核心部位。若X染色体和常染色体间发生染色体易位，使X染色体分裂为两个片段，其中只会有一个片段失活，另一个片段则在所有细胞中保持活性。X染色体上的基因，如果与失活片段相连接就失去表达活性，若与不失活片段相连接则保持表达活性。这就是说，在失活的片段上可能存在着一个失活中心，另一片段很可能由于缺乏该核心区域而保持活性。人们将该核心区命名为X染色体失活中心(X-chromosome inactivation center，Xic)，并定位在人Xq13区(正好是Barr氏小体浓缩部位)和小鼠的3XD。失活的染色体上绝大多数基因都处于关闭状态，DNA序列都呈高度甲基化。

研究人X染色体发现了一个重要的基因 *Xist*(xi-specific transcript)，该基因只在失活的X染色体上表达，而不在活性的X染色体上表达(表8-4)。在小鼠Xic位点上也发现了与人 *Xist* 有高度同源性的 *Xist* 基因。人和小鼠 *Xist* 基因的高度保守性和其染色体定位，以及该基因表达的独特性，暗示它可能参与了X染色体的失活过程。进一步分析发现，人 *Xist* 基因至少有8个外显子，总长约17 kb，其中第一号外显子长达11 364 bp，第六号外显子也长达4 543 bp，整个Xist序列中存在多个串联的重复序列。有趣的是，该基因产物是一功能性RNA分子，不含ORF，却含有大量的终止密码子，因此并不编码蛋白质，而且该RNA分子只存在于细胞核内，不向细胞质转移。实验证明，*Xist* RNA分子能与Xic位点相互作用，引起后者构象变化，更容易与各种蛋白因子相结合，最终导致X染色体失活。

表8-4 *Xist* 甲基化与X染色体失活

| *Xist* | X染色体 | 其他位点 |
|---|---|---|
| 甲基化、不转录 | 活性 | 去甲基化，活性 |
| 去甲基化、转录 | 失活 | 甲基化，失活 |

可以肯定，*Xist* 基因的表达是决定X染色体失活的关键因素。那么，*Xist* 基因的表达调控又是受到何种因素影响的呢？研究发现，活性X染色体上的Xist基因总是甲基化的，而失活X染色体上该位点都是去甲基化的。雄性体细胞只含有一条X染色体，永远处于活性状态，其 *Xist* 基因内部及5′端启动区都高度甲基化。雌性体细胞含有两条X染色体，活性X染色体上 *Xist* 位点都高度甲基化，而失活X染色体上 *Xist* 位点都没有甲基化。

## 8.2 真核基因转录机器的主要组成

真核基因调控主要也是在转录水平上进行的，受大量特定的顺式作用元件(cis-acting element)和反式作用因子(trans-acting factor，又称跨域作用因子)的调控，真核生物的转录调控大多数是通过顺式作用元件和反式作用因子复杂的相互作用来实现的。

## 8.2.1 真核基因的转录

一个完整的基因，不但包括编码区(coding region)，还包括 5′和 3′端长度不等的特异性序列，它们虽然不编码氨基酸，却在基因表达的过程中起着重要作用。所以，“基因”的分子生物学定义是：产生一条多肽链或功能 RNA 所必需的全部核苷酸序列(图 8-16)。由于真核生物 3 种 RNA 聚合酶中，只有聚合酶Ⅱ能够转录信使 RNA 前体，并在加工成熟后按照三联子密码的原理翻译成蛋白质产物，我们在这里主要讨论 RNA 聚合酶Ⅱ的基因转录及其调控过程。

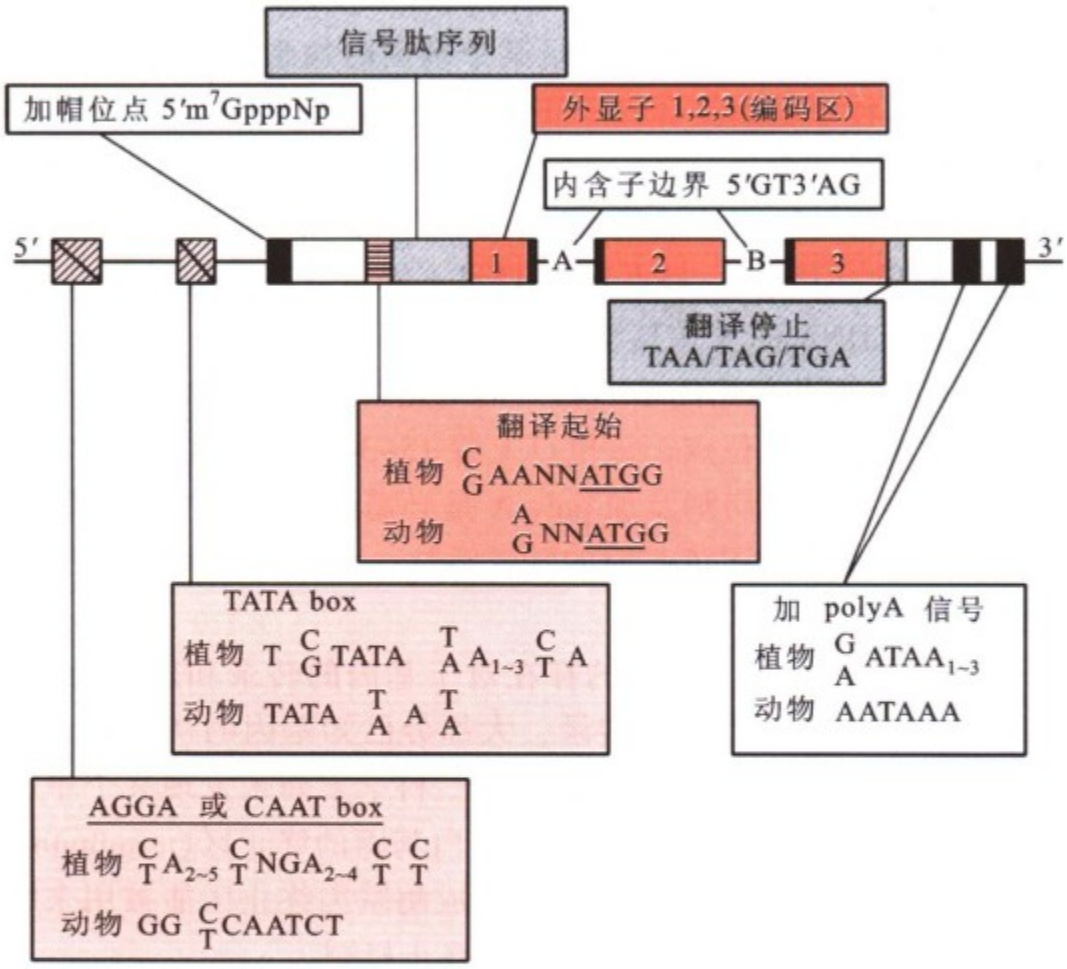



图 8-16 真核基因的一般构造示意图

1. 启动子(promoter)

真核基因启动子由核心启动子和上游启动子两个部分组成，是在基因转录起始位点 (+1) 及其 5′上游大约 100~200 bp 以内的一组具有独立功能的 DNA 序列，每个元件长度约为 7~20 bp，是决定 RNA 聚合酶Ⅱ转录起始点和转录频率的关键元件。

① 核心启动子 (core promoter) 指保证 RNA 聚合酶Ⅱ转录正常起始所必需的、最少的 DNA 序列，包括转录起始位点及转录起始位点上游-30~-25 bp 处的 TATA 盒。核心启动子单独起作用时，只能确定转录起始位点并产生基础水平的转录。

② 上游启动子元件(upstream promoter element，UPE)包括通常位于-70 bp 附近的 CAAT 盒(CCAAT)和 GC 盒(GGGCGG)等，能通过 TFⅡD 复合物调节转

录起始的频率,提高转录效率。

2. 转录模板

包括从转录起始位点到 RNA 聚合酶Ⅱ转录终止处的全部 DNA 序列。

3. RNA 聚合酶Ⅱ

RNA 聚合酶Ⅱ由至少 10~12 个亚基组成，各亚基的相对分子质量在 $1\times10^4$~$2.4\times10^5$,有些亚基也在聚合酶Ⅰ、Ⅲ中共用。其中 $2.4\times10^5$ 最大亚基的羧基端含有由 7 个氨基酸残基(Tyr-Ser-Pro-Thr-Ser-Pro-Ser)组成的多磷酸化位点重复序列,称为羧基端结构域(CTD)。

4. RNA 聚合酶Ⅱ基础转录所需的蛋白质因子(以"TFⅡ"表示)

在生理条件下,RNA 聚合酶Ⅱ转录某个基因时通常需要与 20 种以上的蛋白质因子结合形成转录起始复合物。TFⅡD、TFⅡB 和 TFⅡF 与 RNA 聚合酶Ⅱ可在启动子上形成最初级复合物,开始转录 mRNA,加入 TFⅡE 和 TFⅡH 后形成完整的转录复合物并转录出长链 RNA,加入 TFⅡA 可进一步提高转录效率。RNA 聚合酶Ⅱ沿模板滑动时,TFⅡD 及 TFⅡA 滞留在转录起始位点上,其他因子随聚合酶向模板 DNA 的 3′端移动。

研究 RNA 聚合酶Ⅱ起始的基因转录的终止位点发现，几乎所有真核基因的 3′端都存在一个 poly A 位点，该位点上游 15~30 bp 处的保守序列 AATAAA 对于初级转录产物的准确切割及加 poly A 是必需的。点突变实验将 AATAAA 变为 AAGAAA，虽然维持了该基因的转录活性，却发现 mRNA 的剪接加工受阻,因而没有功能性 mRNA 产生。

尽管 poly A 位点及 AATAAA 的存在对于基因的转录和成熟意义重大,但 RNA 聚合酶Ⅱ却不在该位点终止转录，大部分已知基因的初级转录产物拥有 poly A 位点下游 0.5~2 kb 序列。到目前为止,科学家尚未发现某个单一位点具有特异的转录终止性能。但是,小鼠 $\beta$-珠蛋白基因的终止区(terminator)能被用来终止腺病毒基因的转录,农杆菌胭脂碱合成酶基因终止区能被用来终止几乎所有的外源基因,说明可能存在共同的转录终止机制。

## 8.2.2 增强子及其对转录的影响

增强子是指能使与它连锁的基因转录频率明显增加的 DNA 序列，最早发现于 SV40 早期基因的上游,有两个长 72 bp 的正向重复序列。病毒、植物、动物和人类正常细胞中都发现有增强子存在。重组实验表明,如果把 SV40 增强子上的两个 72 bp 重复序列同时删除,基因表达的水平会降低很多。但如果把该增强子的一个重复序列放回原处或重组 DNA 的任何位置上,则基因转录正常。如果将人 $\beta$-血红蛋白基因克隆到带有 72 bp 重复序列的 DNA 上，这个基因在体内的表达将提高 200 倍以上，即使 72 bp 序列位于该基因转录起始位点上游 3 kb 或下游 2.5 kb,仍不例外。

作为基因表达的重要调节元件,增强子通常具有下列特性:

① 增强效应十分明显。一般能使基因转录频率增加 10~200 倍,有的可以增加上千倍,如经人巨大细胞病毒增强子增强后的珠蛋白基因表达频率比该基

因正常转录高 600~1 000 倍。

② 增强效应与其位置和取向无关。不论增强子以什么方向排列(5′→3′或3′→5′),甚至与靶基因相距 3 000 bp 或在靶基因下游,均表现出增强效应。

③ 大多为重复序列,一般长约 50 bp,适合与某些蛋白因子结合。其内部常含有一个核心序列:(G)TGGA/TA/TA/T(G),该序列是产生增强效应时所必需的。

④ 其增强效应有严密的组织和细胞特异性,说明增强子只有与特定蛋白质(转录因子)相互作用才能发挥功能。

⑤ 没有基因专一性,可以在不同的基因组合上表现增强效应。

⑥ 许多增强子还受外部信号的调控,如金属硫蛋白基因启动区上游所带的增强子,就可以对环境中的锌、镉浓度做出反应。

那么,增强子的作用原理是什么呢?由于增强子常在电镜下呈现环状结构,其活性与半周 DNA 双螺旋(5 bp)的奇、偶倍数有关,表明它的功能受 DNA 双螺旋空间构象的影响。增强子可能有如下 3 种作用机制:①影响模板附近的 DNA 双螺旋结构,导致 DNA 双螺旋弯折或在反式因子的参与下,以蛋白质之间的相互作用为媒介形成增强子与启动子之间"成环"连接,活化基因转录;②将模板固定在细胞核内特定位置,如连接在核基质上,有利于 DNA 拓扑异构酶改变 DNA 双螺旋结构的张力,促进 RNA 聚合酶Ⅱ在 DNA 链上的结合和滑动;③增强子区可以作为反式作用因子或 RNA 聚合酶Ⅱ进入染色质结构的"入口"。

### 8.2.3 反式作用因子

真核生物启动子和增强子是由若干 DNA 序列元件组成的,由于它们常与特定的功能基因连锁在一起,因此被称为顺式作用元件。这些序列组成基因转录的调控区,影响基因的表达。在转录调控过程中,除了需要调控区外,还需要反式作用因子(图 8-17)。

一般认为,如果某个蛋白是体外转录系统中起始 RNA 合成所必需的,它就是转录复合物的一部分。根据各个蛋白质成分在转录中的作用,能将整个复合物分为 3 部分:

① 参与所有或某些转录阶段的 RNA 聚合酶亚基,不具有基因特异性。

② 与转录的起始或终止有关的辅助因子,不具有基因特异性。

③ 与特异调控序列结合的转录因子。它们中有些被认为是转录复合物的一部分,因为所有或大部分基因的启动区都含有这一特异序列。更多的则是基因或启动子特异性结合调控蛋白,它们是起始某个(类)基因转录所必需的。

在诸多转录因子中被广泛研究的主要有识别 TATA 区的 TFⅡD,识别 CAAT 区的 CTF,识别 GGGCGG 的 SP1 以及识别热激蛋白启动区的 HSF。已知每个细胞中可含有约 60 000 个 SP1,而 CTF 的含量则高达每细胞 300 000 个。在 SV40 早期基因启动子区,从-110~-70 区,有 6 个 GC 区同方向串联排列,基因活跃表达时,6 个区全部与 SP1 结合。在胸腺嘧啶激酶基因启动子上,各有一个 GC 区分别位于 TATA 区和 CAAT 区 5′上游端。研究认为,SP1 与 GC 区结合后,可同时在 DNA 链的不同方向与 CAAT 区结合因子(CTF 或 NF1)及 TATA

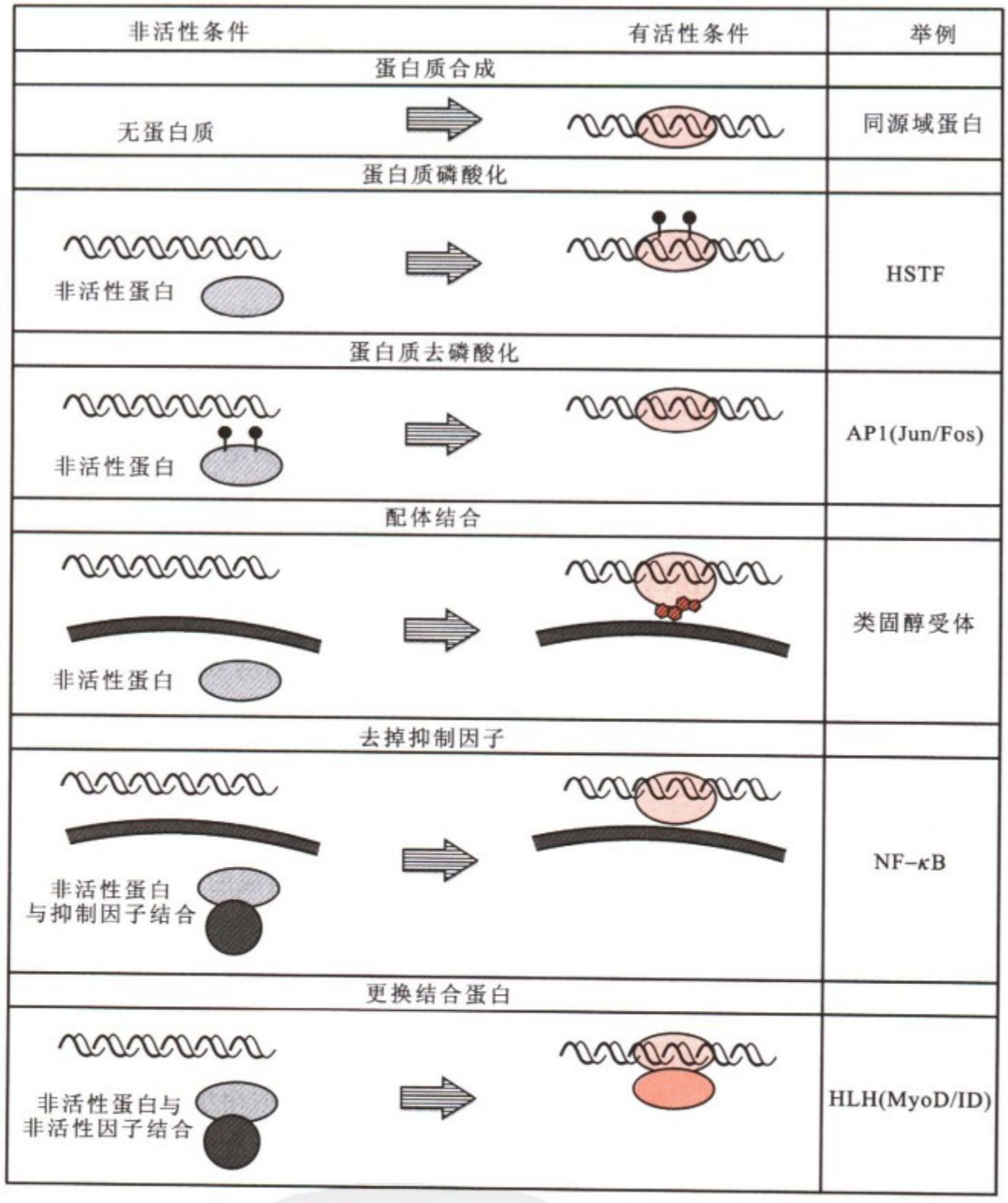



图 8-17 真核生物中转录因子活性调节的主要方式

区结合因子 TFⅡD 发生作用,从而促进 RNA 聚合酶Ⅱ识别特定类型的启动子。

1. DNA 识别或结合域

反式作用因子是能直接或间接地识别或结合在各类顺式作用元件核心序列上(表 8-5),参与调控靶基因转录效率的蛋白质。

(1) 螺旋-转折-螺旋(helix-turn-helix,H-T-H)结构。这一类蛋白质分子中有至少两个 α 螺旋,中间由短侧连氨基酸残基形成“转折”,近羧基端的 α 螺旋中氨基酸残基的替换会影响该蛋白质在 DNA 双螺旋大沟中的结合。控制酵母交配型 *MAT* 基因座以及果蝇体节发育的调节基因 (*antp*、*ftz*、*ubx*) 等同源盒(homeobox)基因所编码的蛋白都有 H-T-H 结构。与 DNA 相互作用时,同源域

表 8-5 部分球蛋白基因 5′调控区序列的保守性比较

| 基因 | -80 区 | -30 区 |
|---|---|---|
| 小鼠 $\beta^{min}$ | GGCC*AAT*CTGCTC | GGT*ATA*TAAA |
| 小鼠 $\beta^{maj}$ | GGCC*AAT*CTGCTC | AGC*ATA*TAAG |
| 兔 β | GGCC*AAT*CTAC -- | GGC*ATA*AAAG |
| 山羊 $\beta^{A}$ | AGCC*AAT*CTGCTC | GGC*ATA*AAAG |
| 山羊 $\beta^{C}$ | AGCC*AAT*CTGCTC | GGC*ATA*AAAG |
| 人 β | GGCC*AAT*CTACTC | GGC*ATA*AAAG |
| 人 δ | AACC*AAT*CTGCTC | -GC*ATA*AAAG |
| 人 $A^{\gamma}$ | GACC*AAT*AGCCTT | GGC*ATA*AAAG |
| 人 $G^{\gamma}$ | GACC*AAT*AGCCTT | GGC*ATA*AAAG |
| 人 α | GACC*AAT*GACTTT | GGC*ATA*AAAG |

注：斜体表示高度保守的序列

蛋白的第一、二两个螺旋往往靠在外侧，其第三个螺旋则与 DNA 大沟结合，并通过其 N 端的多余臂与 DNA 的小沟结合(图 8-18)。

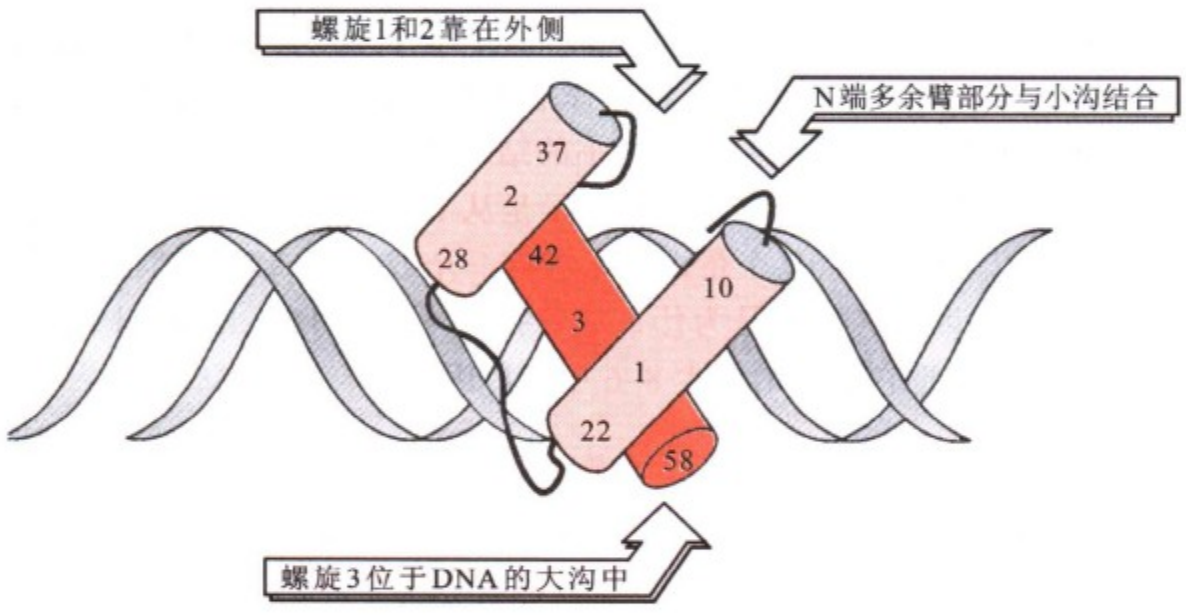



图 8-18 同源域蛋白通过其第三个螺旋与双链 DNA 的大沟相结合，其 N端的延伸部分则与 DNA 的小沟相结合，提高了稳定性

(2) 锌指 (zinc finger) 结构 锌指结构家族蛋白大体可分为锌指、锌钮(twist)和锌簇(cluster)结构，其特有的半胱氨酸和组氨酸残基之间氨基酸残基数基本恒定，有锌参与时才具备转录调控活性。重复的锌指样结构都是一个 α 螺旋与一个反向平行 β 片层的基部以锌原子为中心，通过与一对半胱氨酸和一对组氨酸之间形成配位键相连接，锌指环上突出的赖氨酸、精氨酸参与 DNA 的结合。由于结合在大沟中重复出现的 α 螺旋几乎联成一线，这类蛋白质与 DNA 的结合很牢固，特异性也很高。类固醇激素受体家族含有连续的两个锌指结构，其中两个锌原子将两个 α 螺旋装配成类似 H-T-H 的结构，再以同源或异源性二聚体的方式将两个 α 螺旋结合在相邻的两个大沟中(图 8-19)。

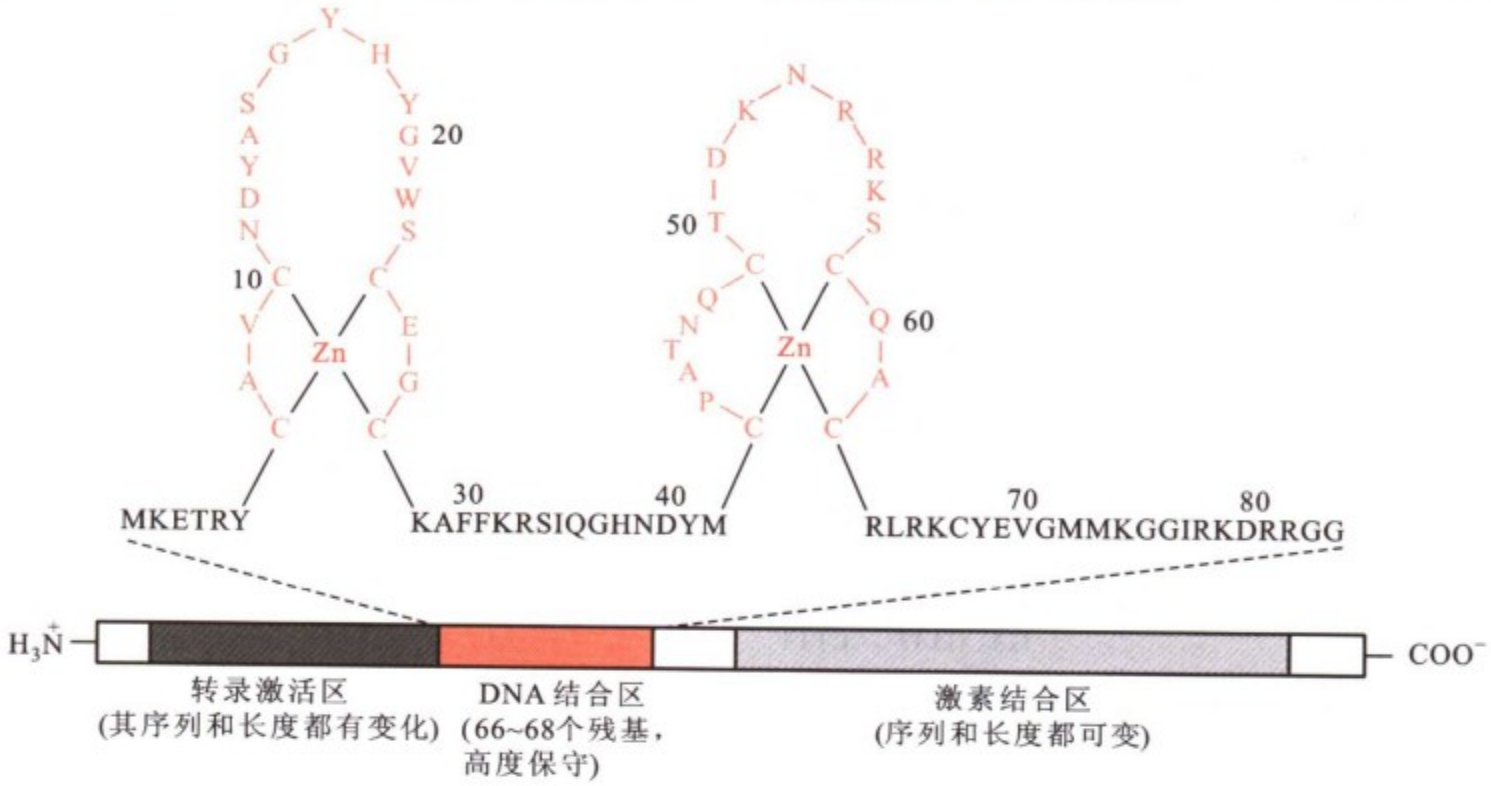



图 8-19 典型的类固醇激素受体结构示意图

大部分蛋白质中的锌指区都聚合成一组，但也发现少量蛋白如果蝇中的 Hunchback 因子,有不止一个锌指簇。锌指簇在这些调控蛋白上的分布也很不相同,在 ADR1 中,锌指区只占一个很小的结构区,而在 TFⅢA 中,整个蛋白几乎被各个锌指区所覆盖。由于锌指区最早是从 RNA 聚合酶Ⅱ和Ⅲ的转录因子中发现的,所以一般认为,某个蛋白如果拥有一个或多个成簇的锌指区,那么它就很可能是转录因子，这就是为什么尽管我们不了解 TDF、Kruppel 及 Hunchback 的功能,但仍把它们归纳在表 8-6 中的原因。

表 8-6 一些转录因子和蛋白质通过 Cys2/His2 与 DNA 相结合

| 蛋白质 | 来源 | 相对分子质量 | 锌指数 | DNA 结合区/bp | 靶序列 | 功能 |
|---|---|---|---|---|---|---|
| TFⅢA | 哺乳动物 | $7.3\times10^4$ | 9 | 50 | *5S* 基因 | 与 RNA 聚合酶Ⅲ作用,促进 5S 基因转录 |
| SP1 | 哺乳动物 | $1.1\times10^4$ | 3 | 10 | GC 区 | 与 RNA 聚合酶Ⅱ作用。广泛性调控因子 |
| ADR1 | 果蝇 | $1.5\times10^4$ | 2 | 22 | *ADH2* 基因 | 与 RNA 聚合酶Ⅱ作用,激活 ADH2 基因 |
| TDF | 哺乳动物 | ? | 13 | ? | ? | 决定雄性性别 |
| Kruppel | 果蝇 | $6.0\times10^4$ | 5 | ? | ? | 胚胎分化 |
| Hunch-Back | 果蝇 | $8.0\times10^4$ | 4+2 | ? | ? | 胚胎分化 |

此外,有些锌指蛋白具有不同的序列:

$$Cys—X_2—Cys—X_{13}—Cys—X_2—Cys$$

这一结构被称为 Cys2/Cys2 锌指，表 8-7 是部分具有这种锌指的转录因子及其一般特征。从表 8-7 中可以看出，这些蛋白一般不具有大量重复性锌指，它们与 DNA 的结合区域短而具有对称性。删除突变分析证实，锌指区是这些蛋白与DNA 相结合所必需的。有人将雌激素受体蛋白中的两个锌指区删去，换上从糖皮质素受体中切出来的两个锌指区后发现，这个融合蛋白只能与糖皮质素受体基因靶序列 GRE 相结合，而不再识别雌激素受体基因本身的靶序列 ERE，说明这类调控蛋白中的锌指区在与靶序列结合的过程中起着主导作用。

表 8-7 具有 Cys2/Cys2 锌指区的转录因子

| 蛋白 | 相对分子质量 | 锌指数 | 靶序列 |
|---|---|---|---|
| 皮质糖受体 | $9.4\times10^4$ | 2 | GRE 中 20 bp |
| 雌激素受体 | $6.6\times10^4$ | 2 | ERE 中 20 bp |
| GAL4 | $9.9\times10^4$ | 1 | $UAS_G$ 中 17 bp |
| 腺病毒 E1A | 约 $3.0\times10^4$ | 1 | ? |

实验表明 Cys2/Cys2 锌指区与 Cys2/His2 锌指区是不同的。如果把固醇类受体蛋白锌指区中的后两个 Cys 突变成为 His，那么靶基因就不会被这个经过改造的转录因子所识别和激活。

(3) 碱性-亮氨酸拉链(basic-leucine zipper)，即 bZIP 结构(图 8-20)。肝脏、小肠上皮、脂肪细胞以及某些脑细胞中存在的一大类 C/EBP 家族蛋白质，它们的特征是能够与 CCAAT 盒和病毒的增强子结合。C/EBP 家族蛋白的羧基端 35 个氨基酸残基具有能形成 α 螺旋的特点，其中每隔 6 个氨基酸就有一个亮氨酸残基，这就导致第 7 个亮氨酸残基都在螺旋的同一方向出现。由于这类蛋白质都以二聚体形式与 DNA 结合，两个蛋白质 α 螺旋上的亮氨酸一侧是形成拉链型二聚体的基础。然而，亮氨酸拉链区并不能直接结合 DNA，只有肽链氨基端 20~30 个富含碱性氨基酸结构域与 DNA 结合。若不形成二聚体，该碱性区对 DNA 的亲和力明显降低。所以，这类蛋白质的 DNA 结合结构域实际是以碱性区和亮氨酸拉链结构域整体作为基础的(图 8-21)。

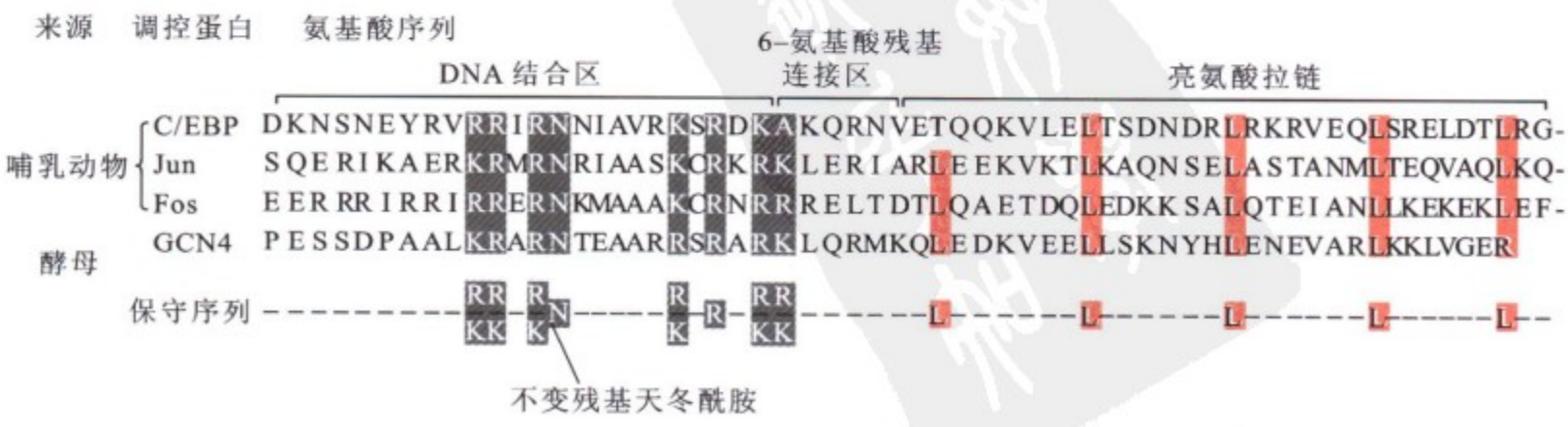



图 8-20 碱性亮氨酸拉链结构域转录激活因子序列比较

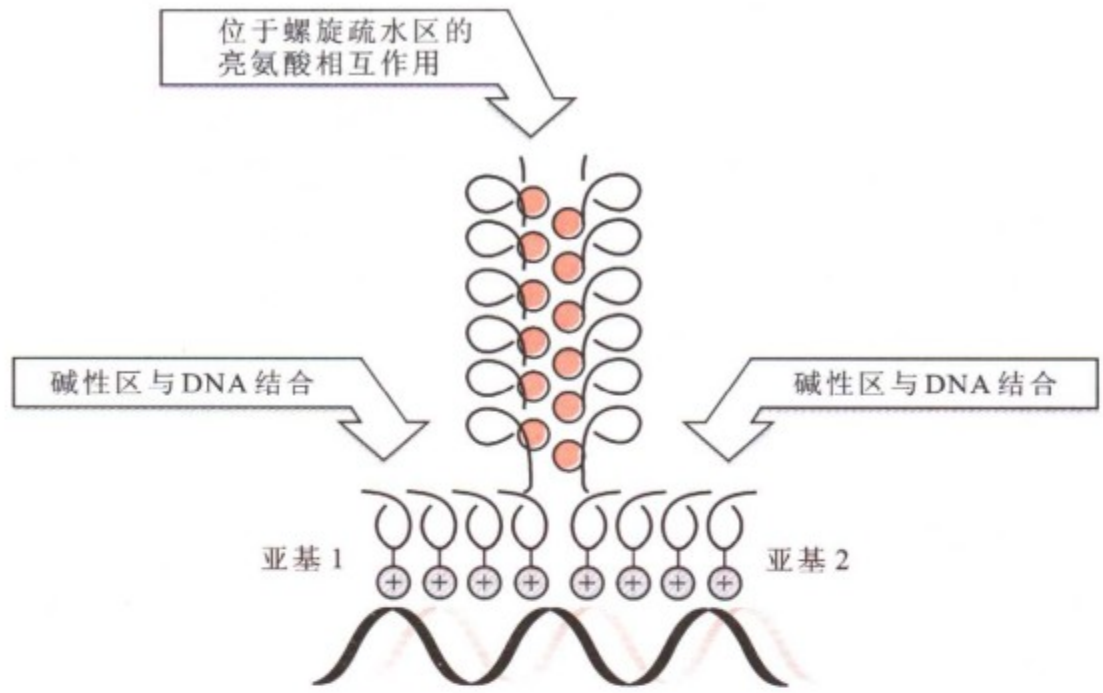



图 8-21 碱性亮氨酸拉链转录激活蛋白与调控区 DNA 结合的示意图

(4) 碱性-螺旋-环-螺旋(basic-helix-loop-helix),即 bHLH 结构。在免疫球蛋白 κ 轻链基因的增强子结合蛋白 E12 与 E47 中,羧基端 100~200 个氨基酸残基可形成两个双性 α 螺旋,被非螺旋的环状结构所隔开,蛋白质的氨基端则是碱性区,其 DNA 结合特性与亮氨酸拉链类蛋白相似。肌细胞定向分化的调控因子 MyoD-1、原癌基因产物 Myc 及其结合蛋白 Max 等都属于 bHLH 蛋白。研究发现,bHLH 类蛋白只有形成同源或异源二聚体时,才具有足够的 DNA 结合能力。当这类异源二聚体中的一方不含有碱性区(如 Id 或 E12 蛋白)时,该二聚体明显缺乏对靶 DNA 的亲和力(图 8-22)。

(5) 同源域(homeo domains)蛋白 同源域是指编码 60 个保守氨基酸序列的 DNA 片段,它广泛存在于真核生物基因组内,由于最早从果蝇 homeotic loci(该遗传位点的基因产物决定了躯体发育)中克隆得到而命名,同源转换基因与生物有机体的生长、发育和分化密切相关。许多含有同源转换区的基因具有转录调控的功能,同源转换区氨基酸序列很可能参与形成了 DNA 结合区。表 8-8 总结了部分含同源转换区转录因子所识别的 DNA 序列,Oct-1、Oct-2 与 Pit-1/GHF-1 所识别的核心序列仅有一个碱基之差,而果蝇 *en*、*ftz* 和 *ubx* 基因产物能识别完全相同的 DNA 序列。*eve* 基因产物除识别与前者相同的序列外,还识别另一个靶序列。

Oct-1 和 Oct-2 专一结合于启动子内 8 碱基区,它们都含有 75 个氨基酸的 pou 区和 60 个氨基酸的同源转换区。尽管 Oct-1 和 Oct-2 中的 homeo box 与经典的果蝇同源转换区变异较大(60 个氨基酸中只有 20 个相同,外加 8 个保守性替换),该区在这两个蛋白质中却是高度保守的(60 个氨基酸中有 53 个相同)。哺乳动物转录因子 Pit-1/GHF-1 及线虫 *unc-86* 基因中都存在类似于 Oct-1 的结构。分析 20 个含同源转换区的果蝇基因发现,该区的保守性高达 80%~90%;即使与亲缘关系很远的哺乳动物同源转换区相比较,同源率也在 35%~40%以

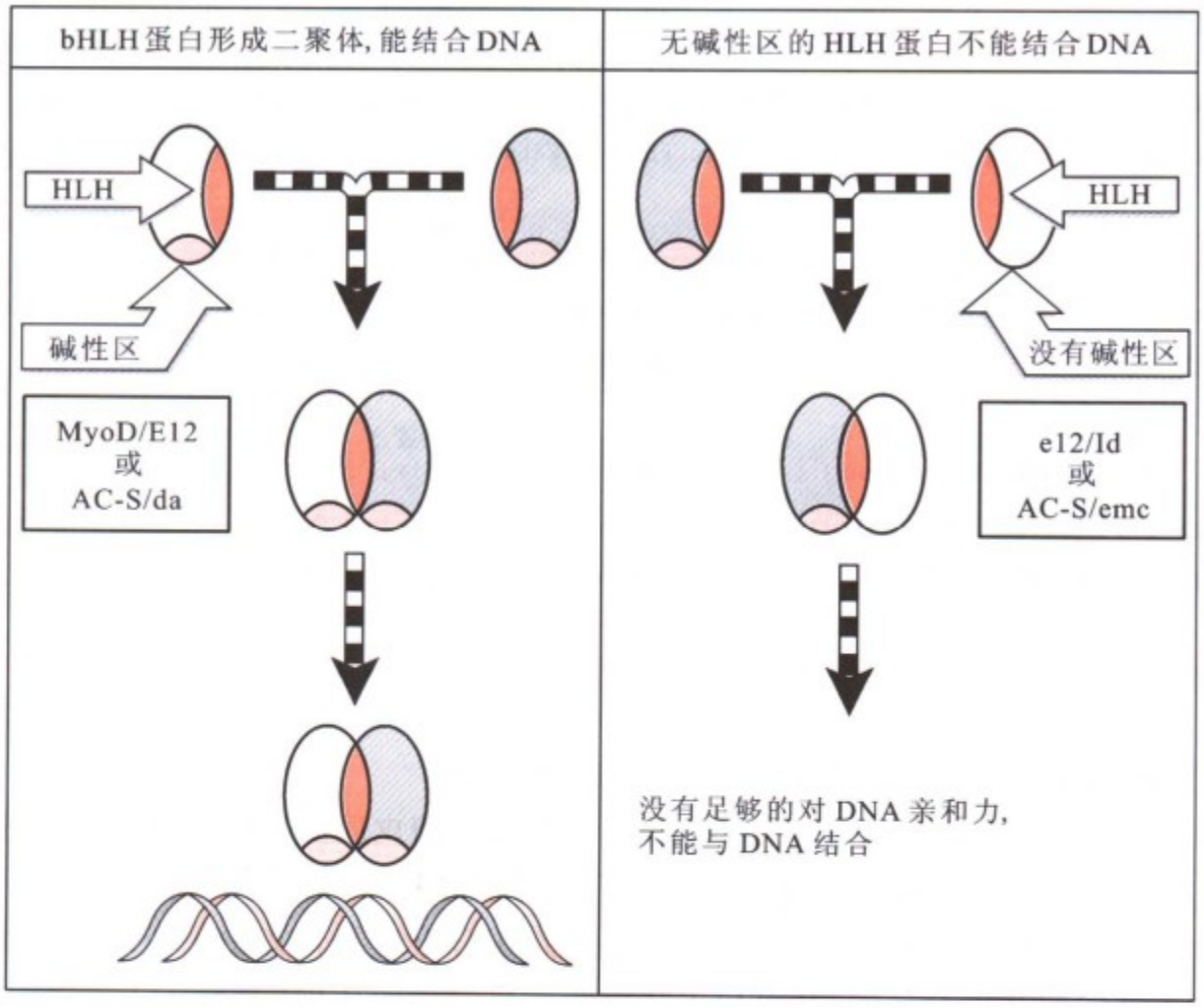



图 8-22 bHLH 蛋白的 DNA 结合能力分析

表 8-8 含有同源转换区的转录因子及其 DNA 靶序列

| 蛋白因子 | 靶 DNA 序列 |
|---|---|
| Oct-1、Oct-2 | ATTTGCAT |
| Pit-1/GHF-1 | TATGCATAA |
| En、ftz、ubx | TCAATTAAAT |
| Eve | TCAATTAAAT TCAGCACCG |
| Bicoid | TCAATCCC |

上(图 8-23)。突变Oct-2 或果蝇 *eve* 和 *ftz* 基因同源转换区某些核苷酸,可以完全防止这些转录因子与 DNA 相互作用。一般认为,同源转换区蛋白之所以具有转录调节功能,与这些蛋白 C 端类似于原核基因阻遏物螺旋-转角-螺旋结构有关。

2. 转录活化结构域

在真核生物中,反式作用因子的功能由于受蛋白质-蛋白质之间相互作用的调节变得精密、复杂,完整的转录调控功能通常以复合物的方式来完成,这就意味着并非每个转录因子都直接与 DNA 结合。因此,是否具有转录活化域就成为反式作用因子中唯一的结构基础。反式作用因子的功能具有多样性,其转录活化域也有多种,通常依赖于 DNA 结合结构域以外的 30~100 个氨基酸残基。

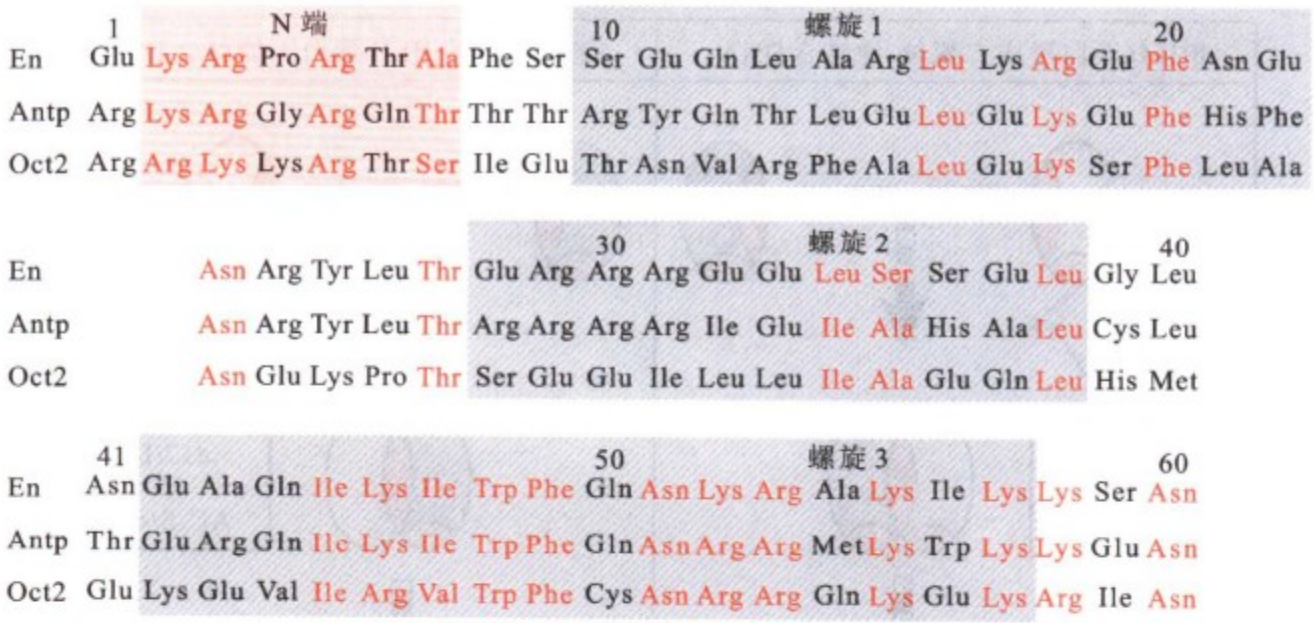



图 8-23 各种同源域 DNA 结合蛋白中保守序列分析

不同的转录活化域大体上有下列几个特征性结构：

(1) 带负电荷的螺旋结构 哺乳动物细胞中糖皮质激素受体的两个转录活化域，AP1 家族的 Jun 及 GAL4 都有酸性的螺旋结构，这种酸性螺旋结构特异性诱导转录起始的活性并不是很强，它们可能与 TFⅡD 复合物中某个通用因子或 RNA 聚合酶Ⅱ本身结合，并有稳定转录起始复合物的作用(图 8-24)。

(2) 富含谷氨酰胺的结构 SP1 是启动子 GC 盒的结合蛋白，除结合 DNA 的锌指结构以外，SP1 共有 4 个参与转录活化的区域，其中最强的转录活化域很少有极性氨基酸，却富含谷氨酰胺，达该区氨基酸总数的 25%左右。哺乳动物细

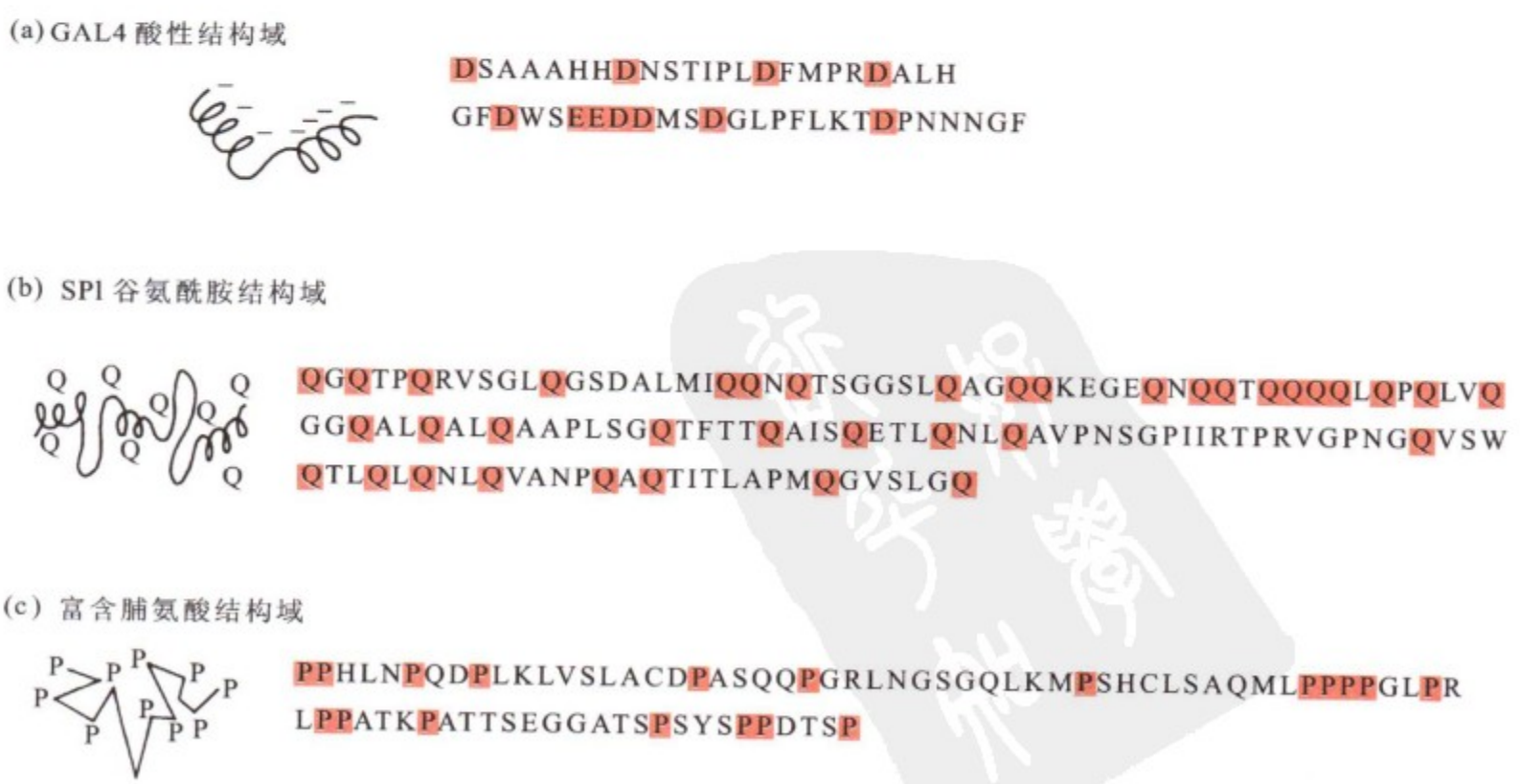



图 8-24 常见的转录活化结构域

胞中的 Oct1/2、Jun、AP2、血清应答因子(SRF)等都有相同的富含谷氨酰胺的结构域。

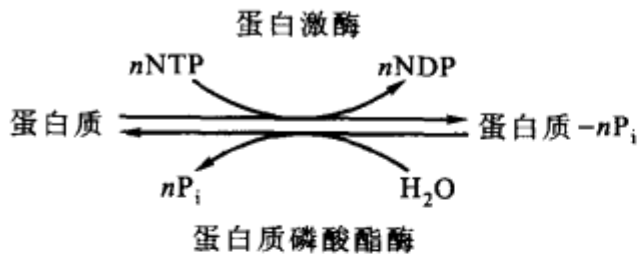
(3) 富含脯氨酸的结构 CTF-NF1 因子的羧基端富含脯氨酸(达 20%~30%),很难形成 α 螺旋。在 Oct2、Jun、AP2、SRF 等哺乳动物因子中也有富含脯氨酸的结构域。

## 8.3 蛋白质磷酸化对基因转录的调控

真核基因转录前基因组中某些位置上的染色质结构改变为伸展型或“开放”状态,为基因的活化奠定了基础并起着“开关”的作用。染色质的伸展结构使细胞核中相应位置上的基因有可能成为 RNA 聚合酶Ⅱ转录的模板,从而启动基因的表达,开始将遗传信息从 DNA 传到 RNA 并翻译为蛋白质,即实现了遗传信息传递的“中心法则”。

细胞是生命活动的基本单位。从低等单细胞生物到复杂的高等生物体内,每个细胞的基因组都携带着各自固有的遗传信息并决定细胞的各种表型特征。细胞除通过 DNA 的复制和随后的细胞分裂将本身所固有的遗传信息由亲代传至子代,实现增殖繁衍之外,它们还不断地“感知”环境(或由于细胞接触带来的)的各种变化,并对其作出特定的应答。细胞应答可以分为 3 个阶段:①外界信息的“感知”,即由细胞膜到细胞核内的信息传递;②染色质水平上的基因活性调控;③特定基因的表达,即从 DNA→RNA→蛋白质的遗传信息传递过程。由此可见,遗传信息的传递无论在细胞的异质性和细胞的特异应答中,都具有关键性的地位,而基因的转录又是遗传信息传递过程中第一个具有高度选择性的环节。因此,对基因转录调控的研究成为近 20 年来分子生物学研究的热点和中心领域。

蛋白质的磷酸化与去磷酸化过程(图 8-25)是生物体内普遍存在的信息传导调节方式,几乎涉及所有的生理及病理过程,如糖代谢、光合作用、细胞的生长发育、神经递质的合成与释放甚至癌变等。




▌图 8-25 蛋白质的磷酸化及去磷酸化过程

细胞表面受体与配体分子的高亲和力特异性结合,能诱导受体蛋白构象变化,使胞外信号顺利通过质膜进入细胞内(图 8-26),或使受体发生寡聚化而被激活。一般情况下,受体分子活化细胞功能的途径有两条:一是受体本身或受体结合蛋白具有内源酪氨酸激酶活性,胞内信号通过酪氨酸激酶途径得到传递;二是配体与细胞表面受体结合,通过 G 蛋白介异的效应系统产生介质,活化丝氨酸/苏氨酸或酪氨酸激酶,从而传递信号。按照“碰撞偶联”或称“动态受体”学说,受体的侧向流动性使受体分子与效应分子在胞质膜磷脂双层内相互作用,这是信号转导的关键。酪氨酸激酶受体侧向扩散使受体寡聚化,是活化信号转导时受体-受体间自身磷酸化所必需的。对于依赖 G 蛋白介导的配体信号转导

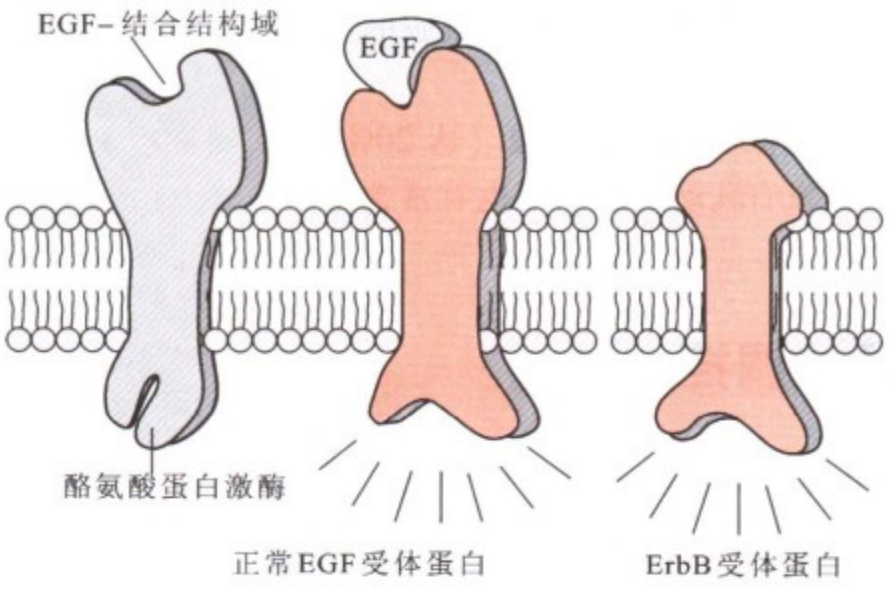



**图 8-26 酪氨酸受体蛋白激酶磷酸化与细胞癌变示意图** 酪氨酸受体蛋白激酶与表皮生长因子(EGF)相结合后,刺激了该受体蛋白的激酶活性,引发一系列生理反应。原癌蛋白 ErbB 虽然没有正常酪氨酸受体蛋白激酶的胞外结构域,其胞内结构域却具有蛋白激酶活性,刺激细胞持久分裂,诱发癌变。

途径来说,配体与受体结合后,G 蛋白和受体-配体复合物由于各自在细胞膜内侧向扩散而发生瞬时随机碰撞导致G 蛋白活化效应分子。此外,受配体结合调控的离子通道也可作为膜上的受体之一参与信息传导(图 8-27)。研究认为,信号转导是维系外部刺激与细胞反应的桥梁, 千变万化的环境因子使信息传导的途径与

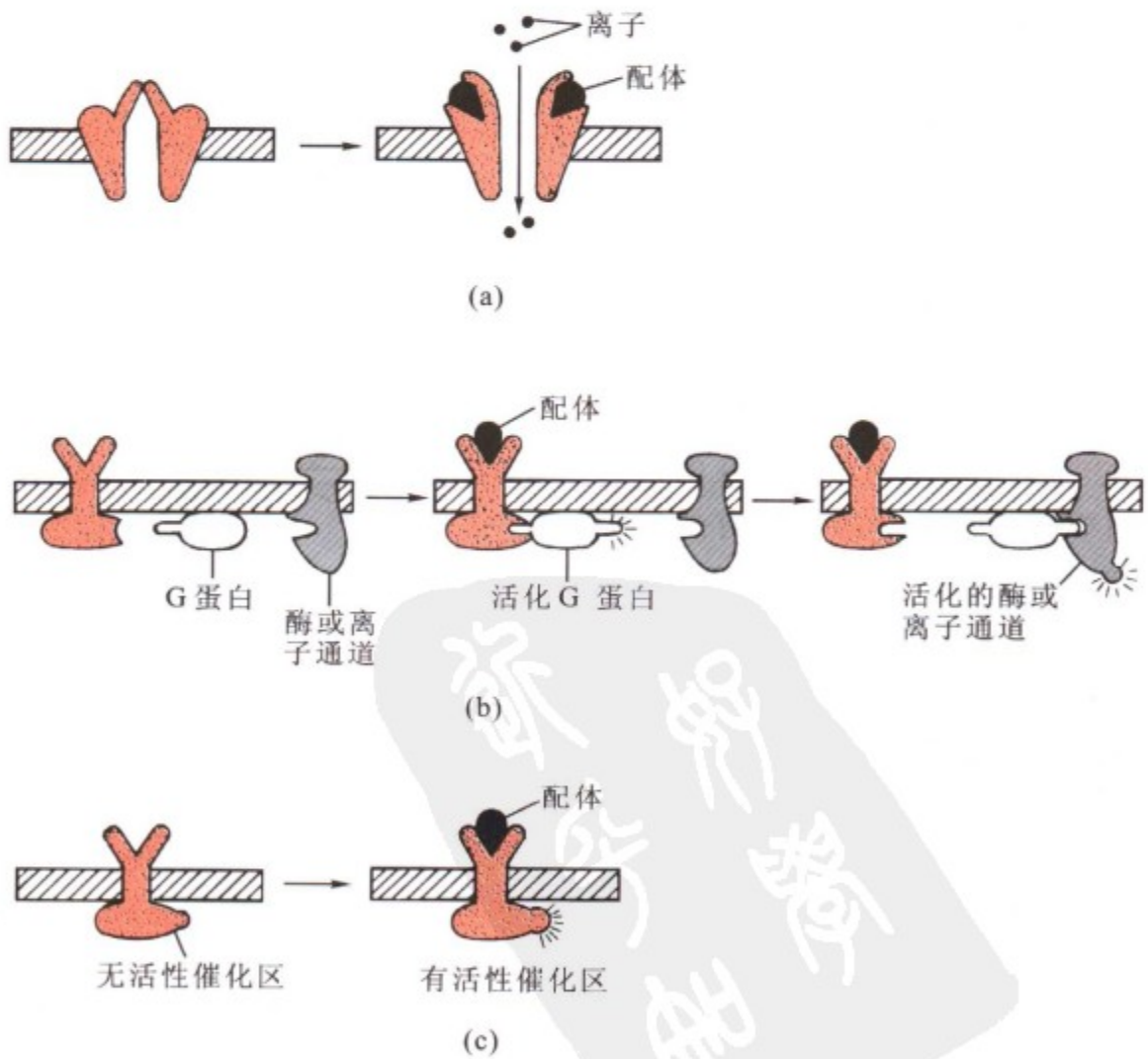



**图 8-27 细胞表面的三类主要受体示意图** (a)受配体结合调控的离子通道;(b)与 G 蛋白相偶联的受体系统;(c)跨膜蛋白激酶-受体系统。

机制成了变幻莫测的迷宫。近年来,数条跨膜信号转导途径已经被逐步阐明(表8-9)。在这些已知的上游途径中,又以酪氨酸蛋白激酶(protein tyrosine kinase,PTK)途径及受体偶联的G蛋白途径较为引人注目(图8-28)。

表8-9 真核细胞主要跨膜信号转导途径

| 体 系 | 作用方式 |
|---|---|
| 蛋白激酶受体 | 配体-受体结合激活受体细胞质内酪氨酸激酶或丝氨酸/苏氨酸激酶 |
| 受体偶联的G蛋白 | 配体结合后激活特定的G蛋白,活化其α亚单位,并调节特定靶分子,如腺苷酸环化酶、磷脂酶、cGMP磷酸二酯酶和离子通道的活性 |
| 离子通道 | 配体结合特定可关闭或开启的离子通道 |
| 心钠素(ANF)受体 | 配体结合活化鸟苷酸环化酶 |
| 细胞因子受体 | 细胞因子与受体结合,抑制或激活核基因表达 |
| 类固醇激素基因家族的核受体 | 与配体结合诱导或抑制靶基因表达 |

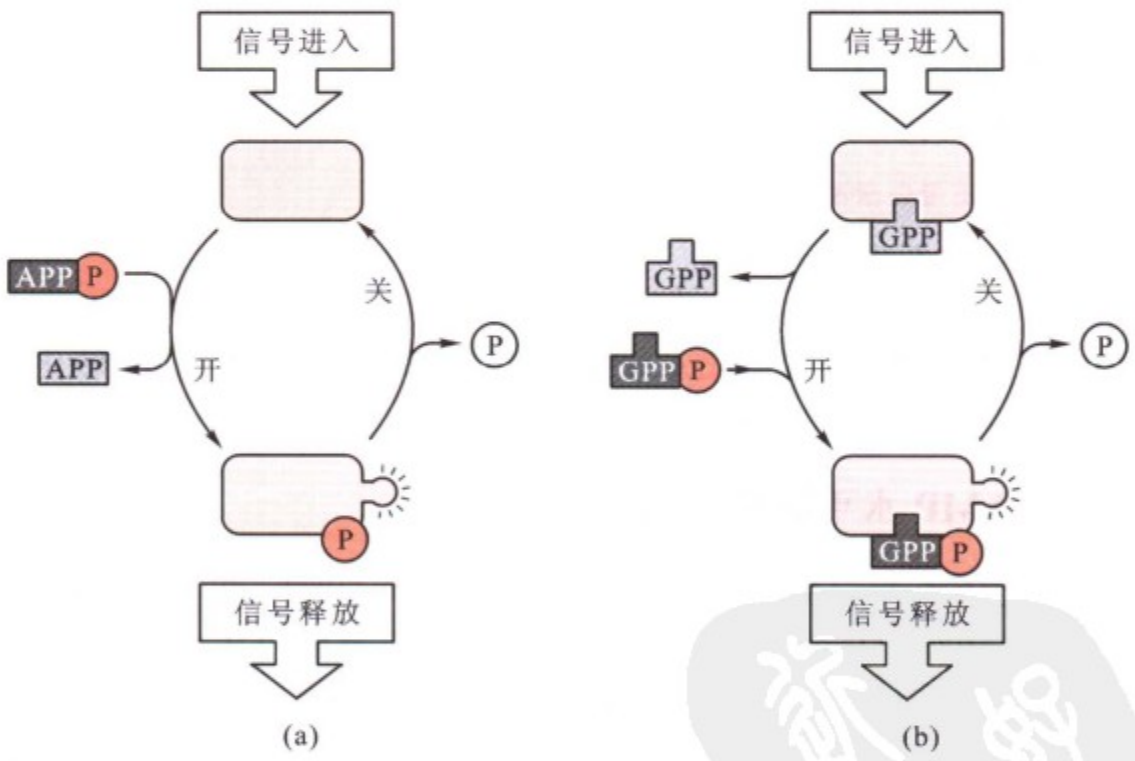



图8-28 蛋白质磷酸化和GTP结合蛋白参与的信号转导过程 (a)蛋白质磷酸化介导的信号转导;(b)GTP结合蛋白介导的信号转导。

到目前为止,已发现的蛋白激酶(protein kinase,PK)基因多达2 000余个,还有约1 000个蛋白质去磷酸化酶基因。根据底物蛋白质被磷酸化的氨基酸残基种类可分为3大类:①丝氨酸/苏氨酸型;②酪氨酸型;③组氨酸型。根据是否有调节物参与蛋白激酶活性可分为两大类:信使依赖型(又分胞内信使、调节因子依赖型和激素或生长因子依赖型,表8-10)和非信使依赖型(表8-11)。细胞

表 8-10 信使依赖型蛋白激酶

| 胞内信使或调节因子依赖型 PK | | | | 激素或生长因子依赖型 PK | | | |
|---|---|---|---|---|---|---|---|
| 名称 | 调节物 | 动物 | 植物 | 名称 | 调节物 | 动物 | 植物 |
| cAMP 依赖型 PK(PKA) | cAMP | + | ? | 表皮生长因子(EGF)依赖型 PK | EGF | + | - |
| cAMP 依赖型 PK(PKG) | cGMP | + | ? | 胰岛素依赖型 PK(胰岛素样生长因子) | 胰岛素 | + | - |
| $Ca^{2+}$、CaM 依赖型 PK | $Ca^{2+}$、CaM | + | + | 血小板衍生生长因子(PDGF)依赖型 PK | PDGF | + | - |
| $Ca^{2+}$、磷脂依赖型 PK(PKC) | $Ca^{2+}$、磷脂 | + | + | 类胰岛素生长因子-1(IGF-1)依赖型 PK | IGF-1 | + | - |
| 只依赖于 $Ca^{2+}$ 的 PK | $Ca^{2+}$ | + | + | 集落刺激因子-1(CSF-1)依赖型 PK | CSF-1 | + | - |

表 8-11 非信使依赖型蛋白激酶

| 名称 | 动物 | 植物 |
|---|---|---|
| 酪氨酸激酶(Ⅰ,Ⅱ型) | + | + |
| 糖原合成酶激酶 | + | ? |
| 组氨酸激酶Ⅱ | + | + |
| 视紫红质蛋白激酶 | + | - |
| β-肾上腺素受体蛋白激酶 | + | - |

受刺激以后,通过蛋白质磷酸化及一系列级联放大过程将胞外信号转化为细胞内信号,从而引起广泛的生理反应。

### 8.3.1 受 cAMP 水平调控的 A 激酶

科学家把依赖于 cAMP 的蛋白激酶称为 A 激酶(PKA),它能把 ATP 分子上的末端磷酸基团加到某个特定蛋白质的丝氨酸或苏氨酸残基上。被 A 激酶磷酸化的氨基酸 N 端上游往往存在两个或两个以上碱性氨基酸,特定氨基酸的磷酸化(X-Arg-Arg-X-Ser-X)改变了这一蛋白的酶活性,动物细胞中这一酶活性代表了 cAMP 所引起的全部反应。在不同的细胞体系中,A 激酶的反应底物是不一样的,这就决定了 cAMP 能在不同靶细胞中诱发不同的反应,非活性状态的 PKA 全酶由 4 个亚基 $R_2C_2$ 所组成,相对分子质量约为 150~170,调节亚基与 cAMP 结合,引起构象变化并释放催化亚基,后者随即成为有催化活性的单体(图 8-29)。研究糖原代谢时第一次发现 cAMP 介导的蛋白质磷酸化过程,激素与其受体在肌肉细胞外表面相结合,诱发细胞质 cAMP 的合成并活化 A 激酶,后者再将活化磷酸基团传递给无活性的磷酸化酶激酶,活化糖原磷酸化酶,最终将糖原磷酸化,进入糖酵解过程并提供 ATP(图 8-30)。

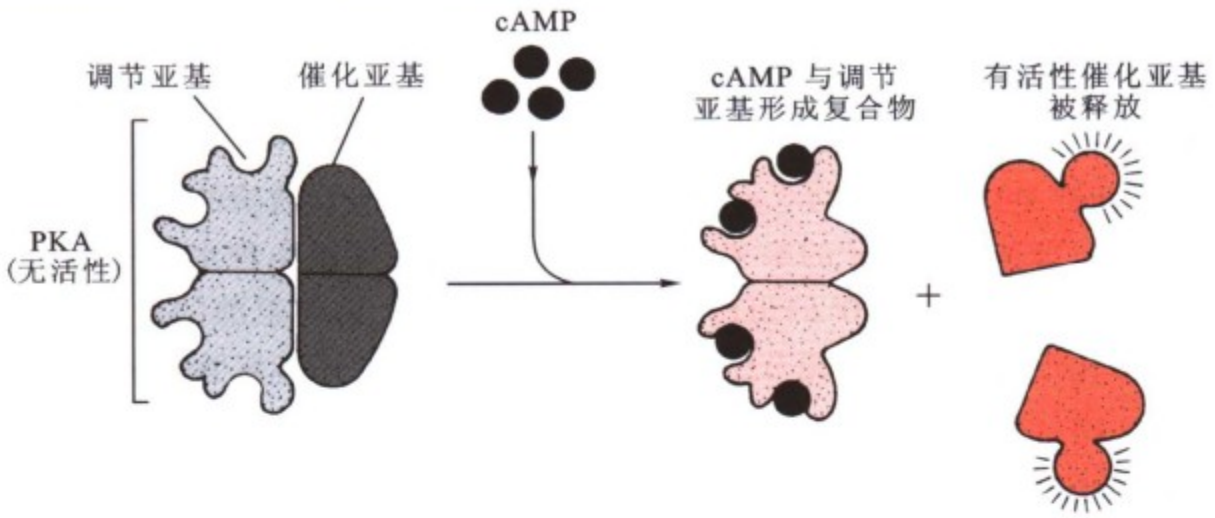



图 8-29 cAMP 调控的 A 激酶活性

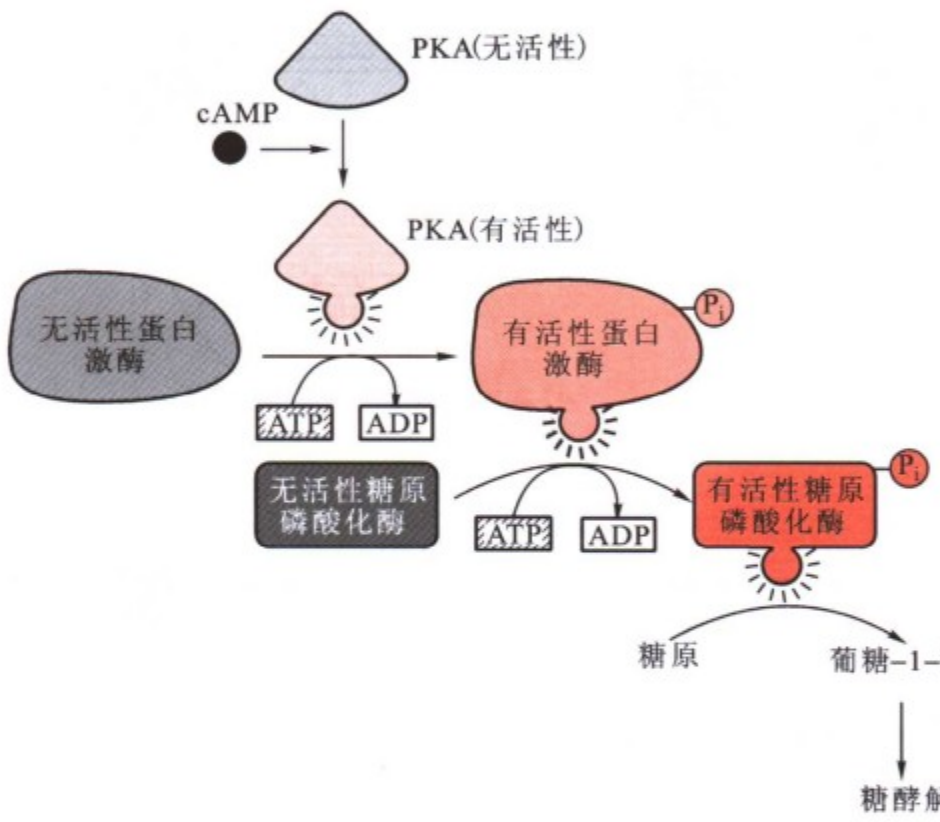



图 8-30 cAMP 活化糖原磷酸化酶示意图

已经证实,许多转录因子都可以通过 cAMP 介导的蛋白质磷酸化过程而被激活,因为这类基因的 5′端启动区大都拥有一个或数个 cAMP 应答元件(cAMP-response element,CRE),其基本序列为 TGACGTCA。膜上的受体 R 与外源配基(ligand)结合,引起受体构象变化,并与 GTP 结合蛋白结合,R 与 G 偶合激活了与膜相关的腺苷酸环化酶(AC),导致胞内 cAMP 浓度上升,活化 A 激酶,释放催化亚基并进入核内,实施底物磷酸化。被磷酸化的底物,如 CREB、CREM 等,可作为转录激活因子诱发基因转录。

### 8.3.2 C 激酶与 $PIP_2$、$IP_3$ 和 DAG

磷酸肌醇级联放大的细胞内信使是磷脂酰肌醇-4,5-二磷酸($PIP_2$)的两个酶解产物:肌醇 1,4,5-三磷酸($IP_3$)和二酰基甘油(DAG)。$IP_3$ 和 DAG 是该途径的

主要活性分子,G 蛋白通过活化的受体调控磷酸肌醇酶系统的活性。图 8-31 是 C 激酶信号传递示意图,因为该蛋白激酶活性是依赖于 $Ca^{2+}$的,所以称为 C 激酶(PKC)。$IP_3$ 引起细胞质 $Ca^{2+}$浓度升高,导致 C 激酶从胞质转运到靠近原生质膜内侧处,并被 DAG 和 $Ca^{2+}$的双重影响所激活。DAG 激活 C 激酶的原因是前者大大提高了 C 激酶对于 $Ca^{2+}$的亲和力,从而使得 C 激酶能被生理水平的 $Ca^{2+}$所活化。C 激酶是一个 $7.7\times10^4$ 的蛋白质,主要实施对丝氨酸/苏氨酸的磷酸化,它具有一个催化结构域和一个调节结构域。与 DAG 结合后还能解除调节区所造成的抑制作用,提高酶活性。

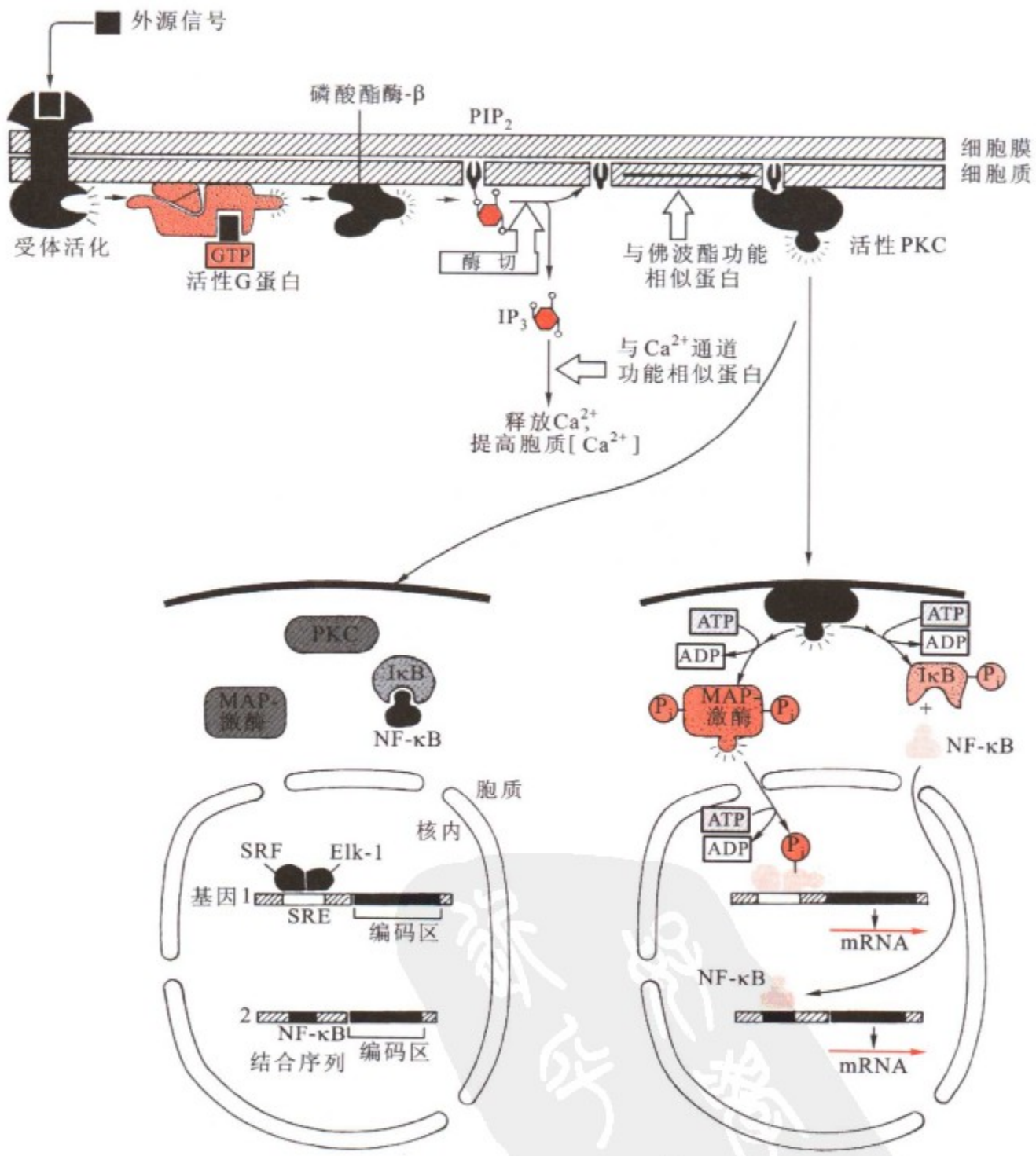



图 8-31 激酶信号传递与基因表达示意图

## 8.3.3 CaM 激酶及 MAP 激酶

$Ca^{2+}$的细胞学功能主要通过钙调蛋白激酶(CaM-kinase)来实现的,它们也

是一类丝氨酸/苏氨酸激酶，但仅应答于细胞内 $Ca^{2+}$ 水平。MAP 激酶(mitogen-activated protein kinase，MAP-kinase，又称为 extracellular-signal-regulated kinase，ERKS)活性受许多外源细胞生长、分化因子的诱导，也受到酪氨酸蛋白激酶及 G 蛋白受体系统的调控。MAP-激酶的活性取决于该蛋白中仅有一个氨基酸之隔的酪氨酸/丝氨酸残基是否都被磷酸化(图 8-32)。科学家把能同时催化这两个氨基酸残基磷酸化的酶称为 MAP-激酶-激酶，它的反应底物是 MAP 激酶。MAP-激酶-激酶本身能被 MAP-激酶-激酶-激酶所磷酸化激活，后者能同时被C 激酶或酪氨酸激酶家族的 Ras 蛋白等激活，从而在信息转导中发挥功能。

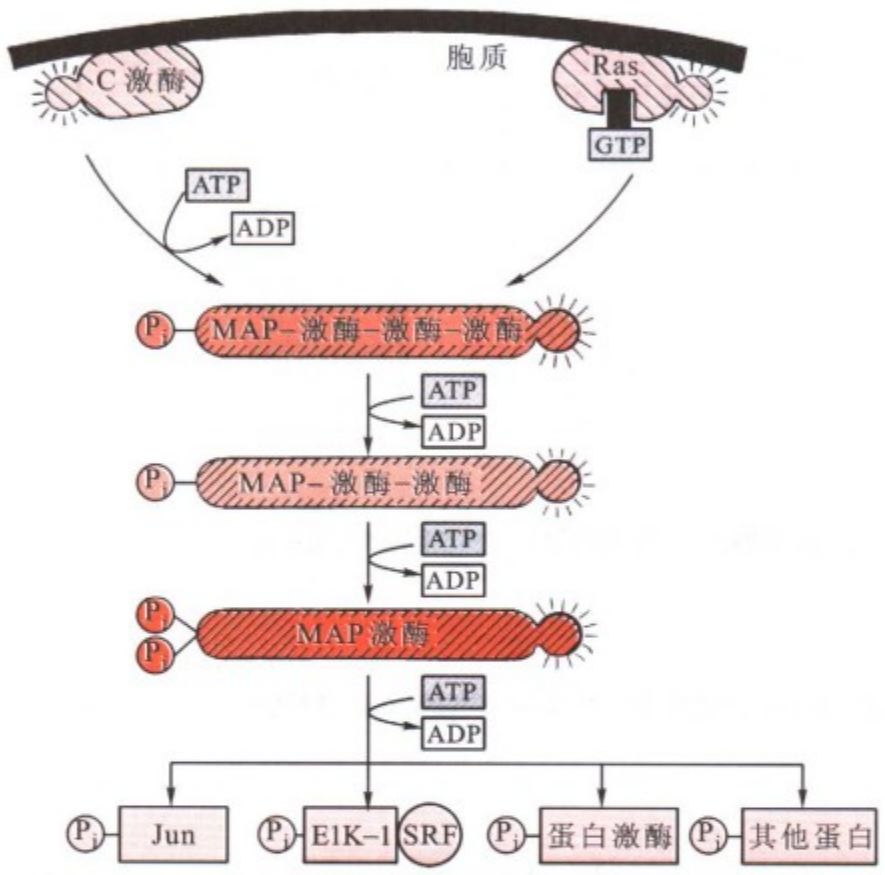


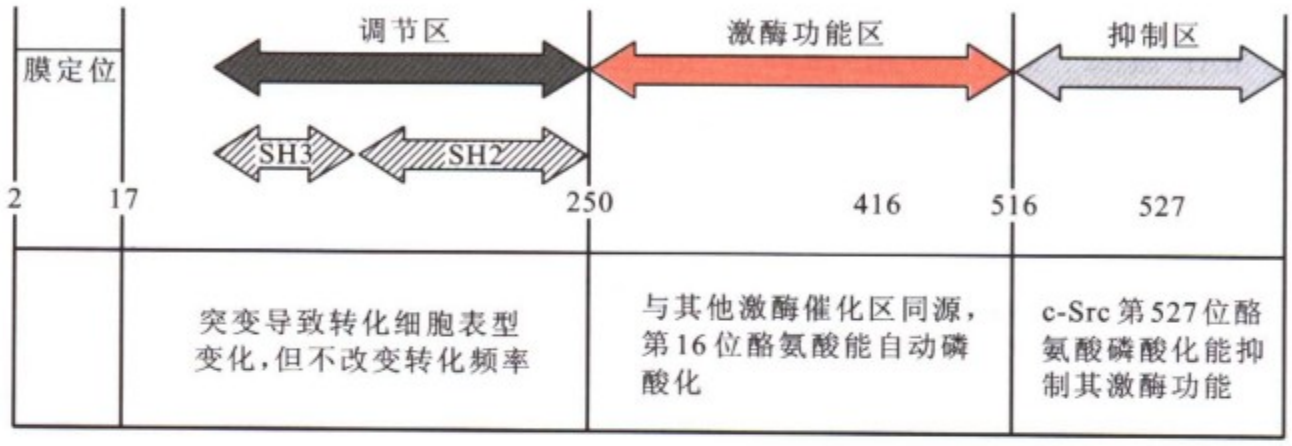

图 8-32 Ras 蛋白和 C 激酶激活丝氨酸/苏氨酸“瀑布式”磷酸化(级联磷酸化)，引起相关生理反应图示

## 8.3.4 酪氨酸激酶途径

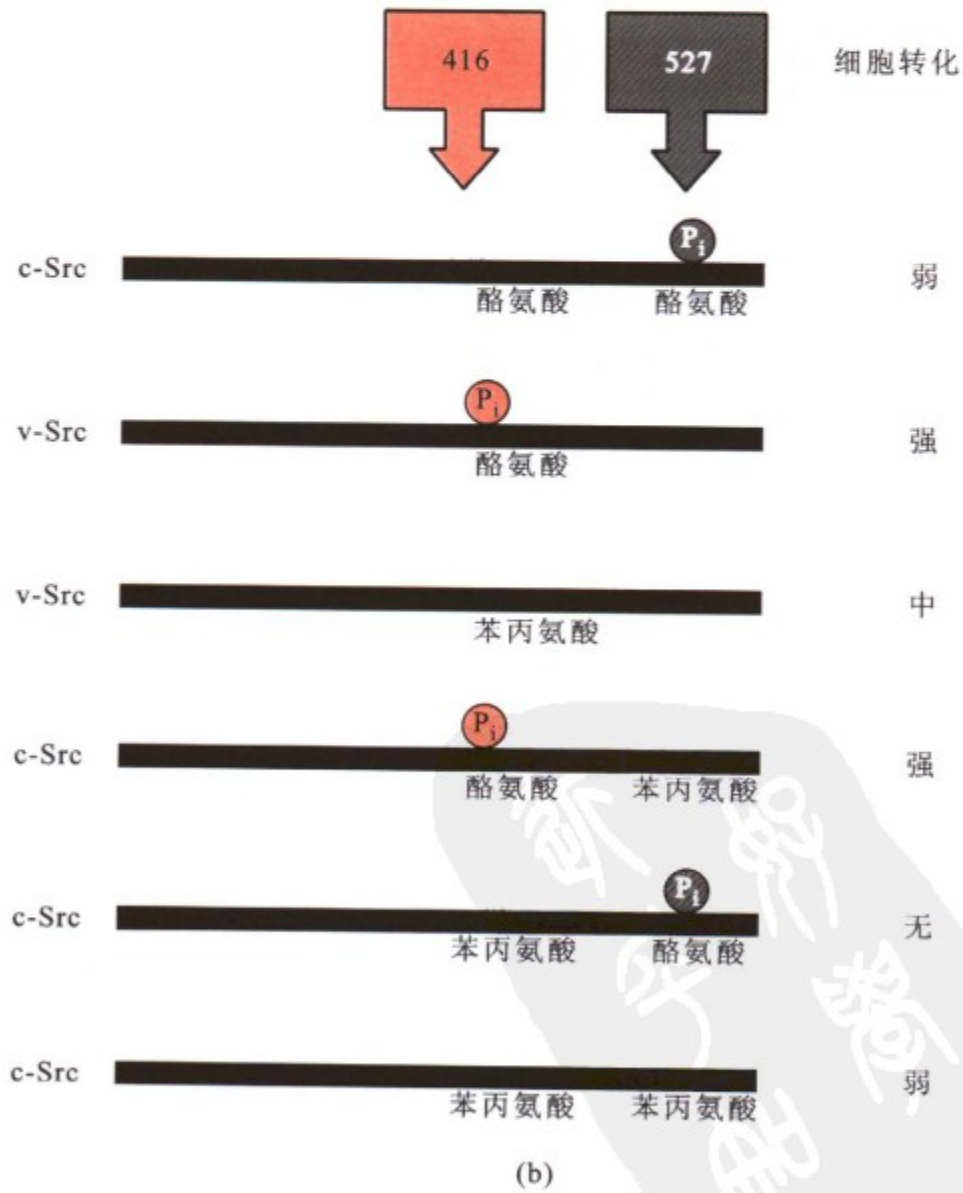
PTK 包含跨膜受体家族与胞质非受体家族两大类。跨膜酪氨酸激酶由胞外结合配体结构域、跨膜结构域和细胞质激酶结构域组成，其 PTK 活性受胞外结构域与配体的调节。配体与受体结合可诱导受体蛋白的二聚化，将受体胞质区酪氨酸残基磷酸化。非受体酪氨酸激酶除有一段与前者同源的激酶结构域(约 300 个氨基酸)序列外，还有数个前者所不具备的保守区。因为 Src 亚家族非受体酪氨酸激酶以及许多 PTK 作用底物分子都存在被称为 Src 同源结构域(Src homology domain，SH)的序列，所以又将激酶功能区称为 SH1，而将另外两个区分别命名为 SH2 和 SH3。SH2 由约 100 个氨基酸残基组成，能与带有磷酸酪氨

酸的蛋白质结合,其特异性由该氨基酸周围的残基序列决定。SH3 由约 60 个氨基酸残基组成，参与将前者定位在质膜内侧或与细胞骨架蛋白的相互作用,与 SH2 结合无关(图 8-33a)。

Src 亚家族成员 $pp60^{v\text{-}Src}$ 定位于质膜内表面，比较转化型 $pp60^{v\text{-}Src}$ 和非转化型 $pp60^{c\text{-}Src}$ 发现，体内 $pp60^{c\text{-}Src}$ 靠近 C 端第 527 位酪氨酸残基高度磷酸化,而




(a)




(b)

图 8-33 c-Src 蛋白功能区划分(a)及第 416 位、527 位酪氨酸残基磷酸化或点突变以后对该蛋白质功能的影响(b)

$pp60^{v\text{-}Src}$ 中这一位点没有被磷酸化。有实验将 $pp60^{c\text{-}Src}Tyr^{527}$ 去磷酸化或使 Tyr→Phe，都能活化 $pp60^{c\text{-}Src}$，显著提高其 PTK 活性和细胞转化能力(图 8-33b)。科学家认为，在正常细胞中，$pp60^{c\text{-}Src}$ 上的 $Tyr^{527}$ 被磷酸化并以头尾相连的缔合方式折入同一分子的 SH2 结构域中，从而抑制了 PTK 活性。$pp60^{c\text{-}Src}$ 与靶分子 $PI_3$ 激酶，rasGAP 等的缔合作用均涉及 $pp60^{c\text{-}Src}$SH2 结构域与 $PI_3$ 激酶及 rasGAP 分子上酪氨酸磷酸化区的结合(图 8-34)。

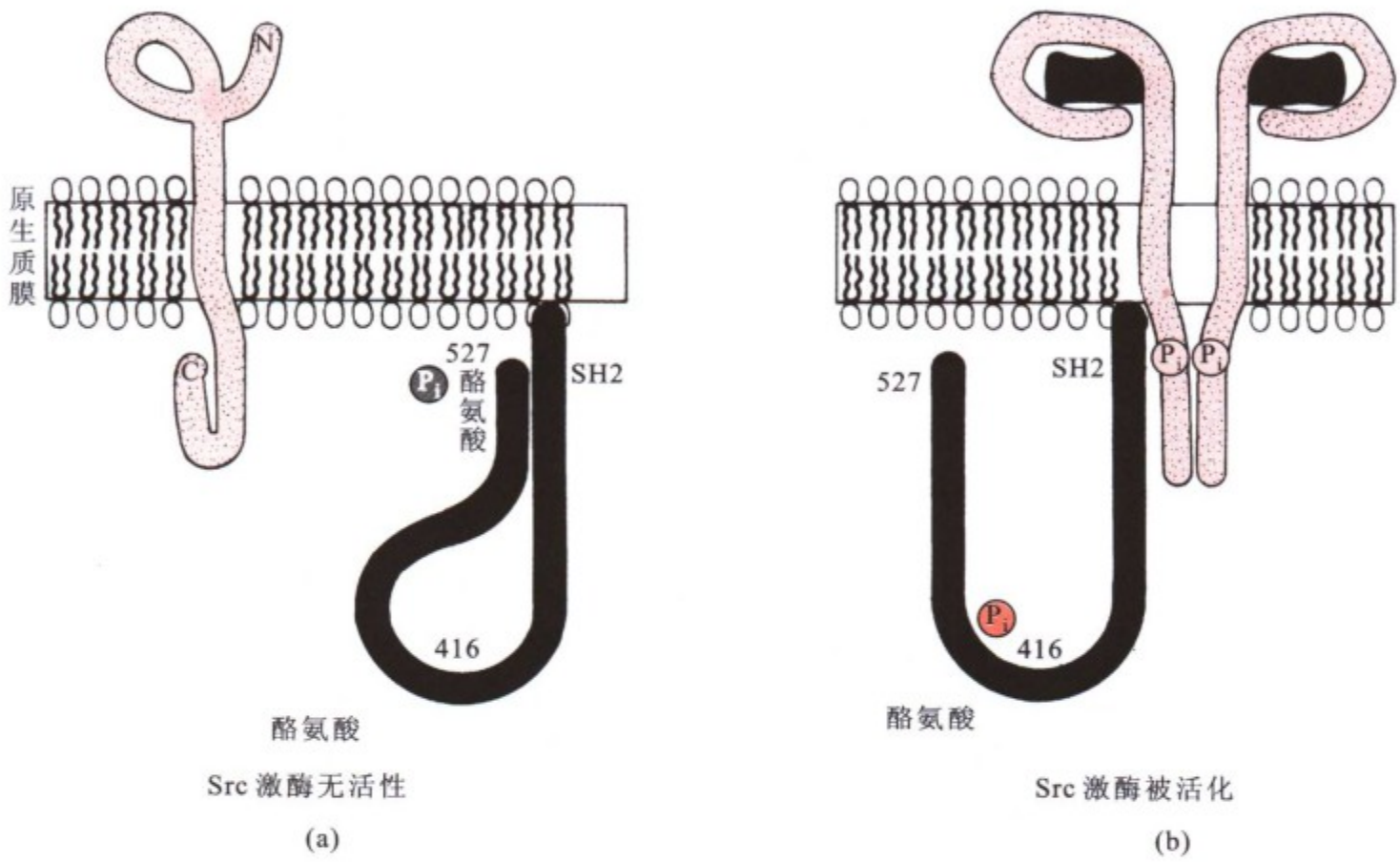



图 8-34 Src 激酶活化过程图示 (a)第 527 位酪氨酸被磷酸化后能与 SH2 区结合，阻碍了受体分子的结合，无酶活性；(b)第 527 位酪氨酸去磷酸化后，SH2 区与膜受体结合，第 416 位酪氨酸被磷酸化，有酶活性。

### 8.3.5 蛋白质磷酸化与细胞分裂调控

细胞通过 p53 及 p21 蛋白控制 CDK(cyclin-dependent protein kinase)活性，调控细胞分裂的进程。如果 p21 蛋白过量，大量细胞周期蛋白(cyclin)E-CDK2 复合物与p21 蛋白相结合，使 CDK2 丧失磷酸化 pRb 蛋白的功能。没有被磷酸化的 pRb 蛋白与转录因子 E2F 结合并使后者不能激活一系列与 DNA 合成有关的酶，导致细胞不能由 G1 期进入 S 期，细胞分裂受阻。如果细胞中 *p53* 基因活性降低，p21 蛋白含量急剧下降，细胞周期蛋白 E-CDK2 复合物就能有效地将 pRb 蛋白磷酸化。此时，pRb 蛋白不能与 E2F 相结合，使后者发挥转录调控因子的作用，激活许多与 DNA 合成有关的基因表达，使细胞从 G1 期进入 S 期，引发细胞分裂(图 8-35)。

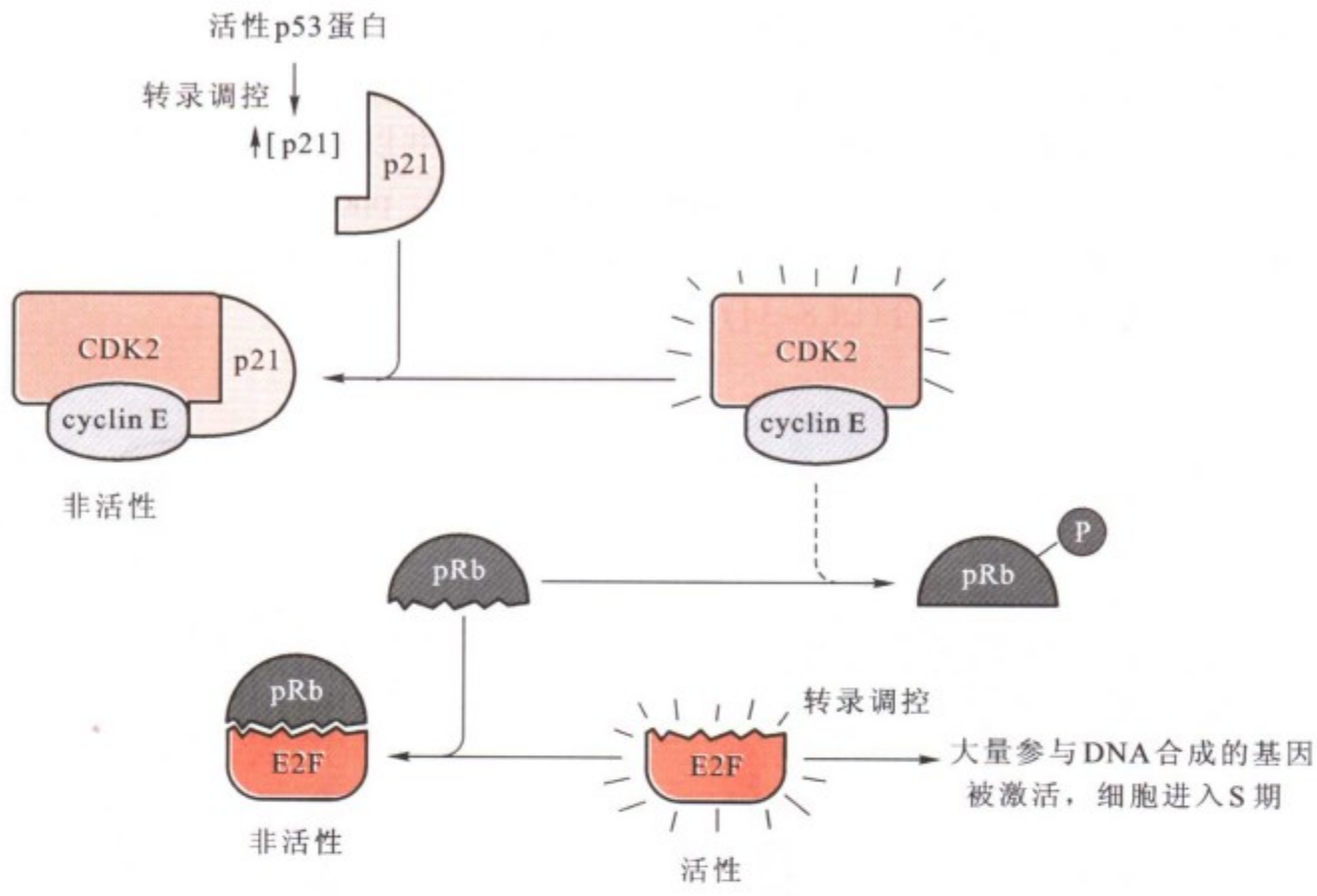



图 8-35 pRb 蛋白的磷酸化和去磷酸化过程控制了细胞分裂的周期性

研究还发现,CDK 的活性也受细胞周期蛋白和磷酸化调控(图 8-36)。当细胞中没有周期蛋白时,CDK 无活性。随着周期蛋白的合成和积累,逐步形成周期蛋白-CDK 复合物,但此时 CDK 蛋白上的酪氨酸位点被优先磷酸化,掩盖了其 ATP 结合位点,ATP 不能有效地与之结合,CDK 仍然无活性。只有当 CDK 蛋白 T-环中的苏氨酸位点被磷酸化,并将其酪氨酸位点上的磷酸基团去掉,CDK 蛋白才得以发挥其生物学活性。同时,CDK 蛋白还能使细胞中的磷酸酯酶磷酸化,以加速脱去自身酪氨酸位点上的磷酸基团。有生物活性的周期蛋白-CDK 复合物也能将 DBRP(destruction box recognizing protein)磷酸化,激活泛蛋白连接酶,把大量泛蛋白加到细胞周期蛋白上并使之迅速降解,CDK 失活,新的周期开始。

## 8.4 蛋白质乙酰化对基因表达的影响

蛋白质乙酰化普遍存在于在生物体细胞内,涉及许多生理生物化学过程。目前主要研究组蛋白乙酰化对基因表达调控的影响。

### 8.4.1 组蛋白的乙酰化及去乙酰化

1. 组蛋白的基本组成

组蛋白(histone)是组成核小体的基本成分,核小体(nucleosome)是组成染色质的基本结构单元。核小体由组蛋白八聚体(由两个包含 $H_2A$,$H_2B$,$H_3$ 和 $H_4$ 的四聚体组成) 和缠绕两圈的 DNA 组成。在相邻的核小体之间有一个 20~

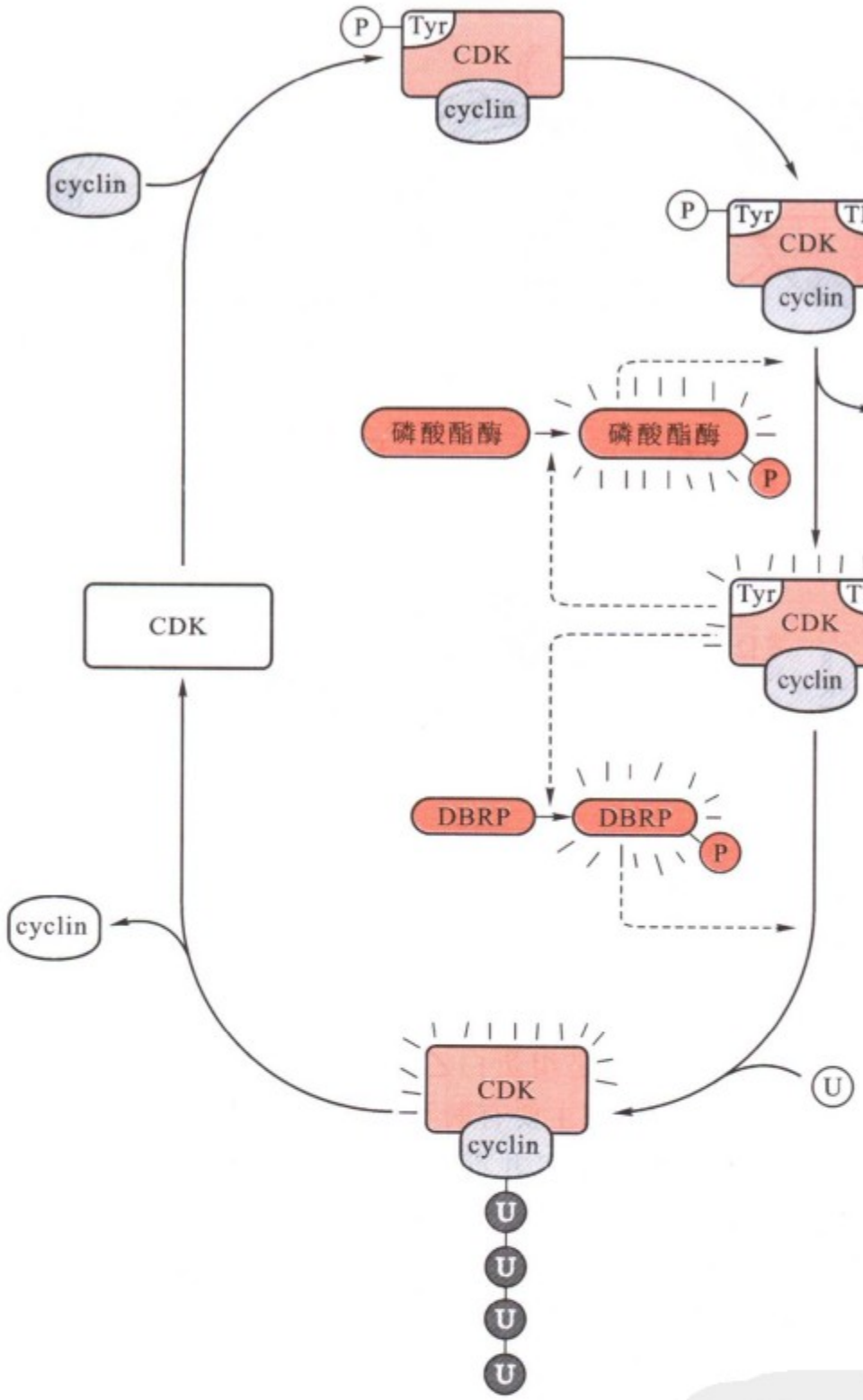



图 8-36 CDK 活性受磷酸化和蛋白质水解通路直接调控

200 bp 的间隔区，在电子显微镜下，一列核小体看上去像一串珠子，每个珠子的直径大约是 10 nm。另一个组蛋白 $H_1$ 结合在核小体核心之外，起到稳定核小体序列和染色质的高级结构的作用。

2. 核心组蛋白的乙酰化和去乙酰化

核心组蛋白朝向外部的 N 端部分被称为“尾巴”，可被组蛋白乙酰基转移酶和乙酰基去乙酰化酶修饰，加上或去掉乙酰基团。图 8-37 是已报道的组蛋白 N 端“尾巴”主要被修饰位点。

3. 组蛋白乙酰基转移酶

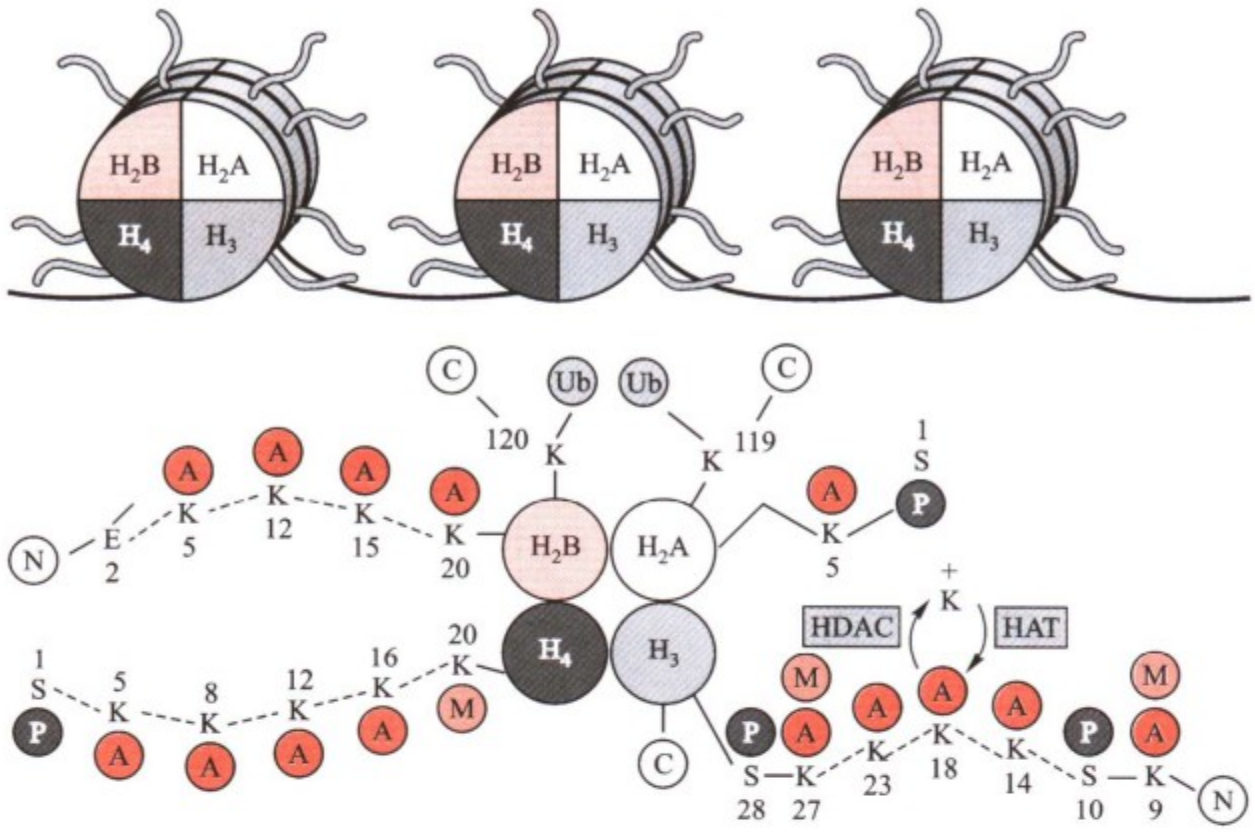



**图 8-37 已知组蛋白 N 端"尾巴"主要被修饰位点** N,组蛋白 N 端;C,组蛋白 C 端。A,乙酰基团;P,磷酸基团;M,甲基基团;Ub,泛蛋白基团。HAT,乙酰基转移酶;HDAC,乙酰基去乙酰化酶。

催化组蛋白乙酰化反应的是组蛋白乙酰转移酶 (histone acetyltransferase, HAT)。目前已经发现的 HAT 酶主要有两类:一类与转录有关,另一类与核小体组装以及染色质的结构有关。HAT 并不是染色质结合蛋白,但是可以通过与其他蛋白相互作用来影响染色质的结构。许多转录激活因子都有 HAT 活性(图 8-38),与转录有关的 HAT 催化亚基是酵母调控蛋白质 GCN5 的同源物,而 GCN5 本身具有乙酰化组蛋白 $H_3$ 和 $H_4$ 的活性。

TAF Ⅱ 250 组蛋白乙酰基转移酶是 TF Ⅱ D 复合物的一个亚基,它的主要功能是与启动子结合协助起始转录,能使组蛋白 $H_3$ 和 $H_4$ 乙酰化。转录共激活子 PCAF 也能使这两个组蛋白乙酰化。组蛋白乙酰基转移酶 p300/CBP 也是转录共激活子,通过与增强子结合蛋白相互作用来调节转录,能使组蛋白 $H_2A$,$H_2B$,$H_3$ 和 $H_4$ 乙酰化。与激素受体协同作用的转录共激活子 ACTR 也是一种作用于 $H_3$ 和 $H_4$ 的组蛋白乙酰基转移酶(表 8-12)。

4. 组蛋白去乙酰化酶

组蛋白去乙酰化酶(histone deacetylase,HDAC)负责去除组蛋白上的乙酰基团,目前研究比较深入的是人类中的 HDAC1 和酵母中的 Rpd3。HDAC 和Rpd3 都形成很大的蛋白复合体发挥作用。Rpd3 能特异性去除组蛋白上的乙酰基团,使核小体相互靠近,并在转录共抑制子 Sin3 及 R 的协同作用下,抑制基因转录(图 8-39)。

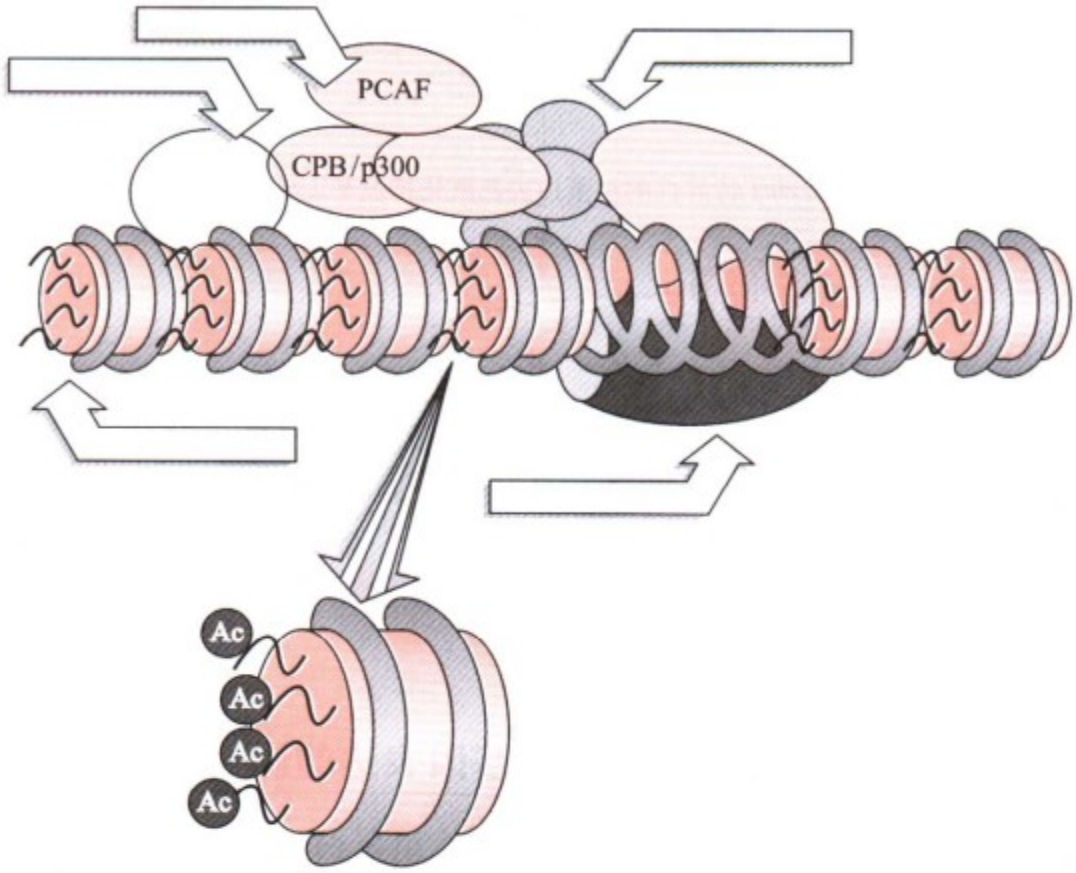



图 8-38 具有组蛋白乙酰基转移酶活性的蛋白参与转录调控示意图

表 8-12 部分已知具有乙酰基转移酶活性的蛋白

| HAT | 是否与转录相关 | 组蛋白体外特异性 |
|---|---|---|
| GNAT 家族 | | |
| Hat1 | 无(与组蛋白组装相关的 B 类 HAT) | $H_4$ |
| Gcn5 | 转录共激活子 | $H_3/H_4$ |
| PCAF | 转录共激活子 | $H_3/H_4$ |
| Hpa2 | 未知 | $H_3/H_4$ |
| MYST 家族 | | |
| Sas3 | 沉默 | $H_3/H_4/H_2A$ |
| Esa1 | 细胞增殖 | $H_4/H_3/H_2A$ |
| MOF | 剂量补偿 | $H_4/H_3/H_2A$ |
| Tip60 | HIV 转移活化基因相互作用 | $H_4/H_3/H_2A$ |
| MORF | 未知 | H4/H3/H2A |
| p300/CBP | 通用型共激活子 | $H_2A/H_2B/H_3/H_4$ |
| SRC-1 | 核受体共激活子,激素信号应答 | $H_3/H_4$ |
| ACTR | 核受体共激活子,激素信号应答 | $H_3/H_4$ |
| TAF Ⅱ 250 | TBP 相关联的因子 | $H_3/H_4$ |
| F Ⅲ C90 | RNA 聚合酶Ⅲ起始转录 | $H_3$ |

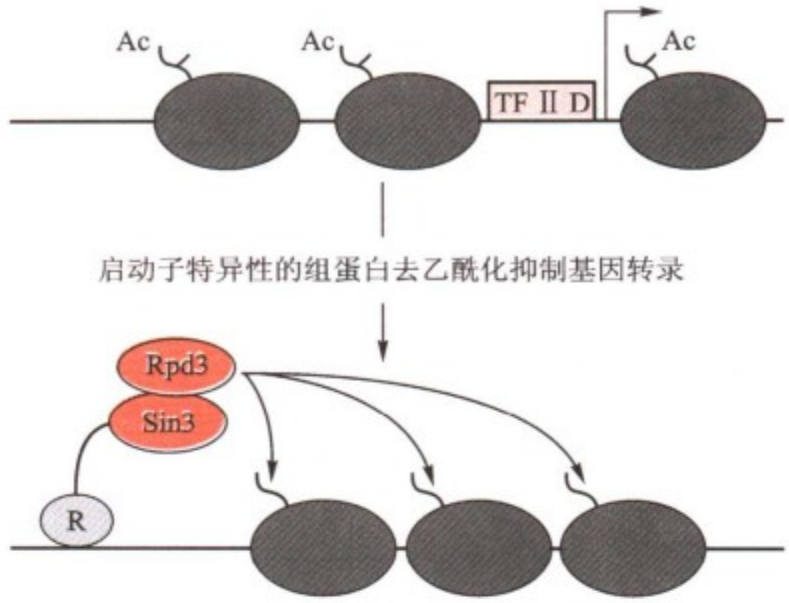



图 8-39 组蛋白去乙酰化酶复合体能与靶基因启动子区结合,抑制基因转录

## 8.4.2 组蛋白乙酰化及去乙酰化对基因表达的影响

组蛋白乙酰转移酶和去乙酰化酶通过使组蛋白乙酰化和去乙酰化对基因表达产生影响。组蛋白 N 端"尾巴"上赖氨酸残基的乙酰化中和了组蛋白尾巴的正电荷,降低了它与 DNA 的亲和性,导致核小体构象发生有利于转录调节蛋白与染色质相结合的变化,从而提高了基因转录的活性。核心组蛋白 $H_2A$,$H_2B$,$H_3$,$H_4$ 通过组蛋白"尾部"选择性乙酰基化影响核小体的浓缩水平和可接近性。研究发现,受乙酰化修饰的组蛋白形成的核小体的结构比未经修饰的组蛋白核小体松散。由于乙酰化的组蛋白抑制了核小体的浓缩,使转录因子更容易与基因组的这一部分相接触,有利于提高基因的转录活性。

组蛋白乙酰基转移酶和去乙酰化酶只能有选择地影响一部分基因的转录。哺乳动物和酵母中得到的实验数据都表明,转录抑制主要是由于新产生的 HDAC/Rpd3 复合体专一性结合于某个或某类基因启动子区附近的组蛋白位点并使之去乙酰化,导致染色质结构发生不利于基因转录的变化。

## 8.4.3 p53 乙酰化对转录活性的影响

肿瘤抑制因子 p53 蛋白与人类多种肿瘤有着密切的关系,因此可能是该领域中研究最为深入的蛋白质。p53 参与多种信号通路,如细胞周期调控、DNA 损伤修复、血管的生成与抑制以及细胞凋亡等的调控。已知 p53 蛋白的活性受翻译后修饰(如磷酸化、乙酰化等)机制调控,经修饰的 p53 蛋白能与不同的靶分子或蛋白复合体结合,从而抑制或激活参与特定生理反应的靶基因。p53 在所有细胞中都转录产生 2.2~2.5 千核苷酸的 mRNA,编码由 393 个氨基酸组成的相对分子质量为 53 000 的蛋白(p53 蛋白)。根据 p53 的结构特点可将其分为 3 个不同区域:N 端酸性区(1~80 位氨基酸残基);C 端碱性区(319~393 位)和中间(100~300 位)的疏水区(图 8-40a)。

p53 能识别不同构象靶基因的相同序列。靶基因启动子的拓扑结构和 p53 蛋白(及与其相互作用的辅助蛋白)的构象决定该蛋白是否与启动子区相互作用。乙酰化使 p53 蛋白的 DNA 结合区域暴露,增强了 DNA 结合能力,从而促进了靶基因的转录(图 8–40b)。CBP/p300 等蛋白复合体既能诱导染色体结构发生有利于结合 p53 蛋白的改变,又使 p53 蛋白被乙酰化,从而显著提高受 p53 调控的靶基因的转录活性。

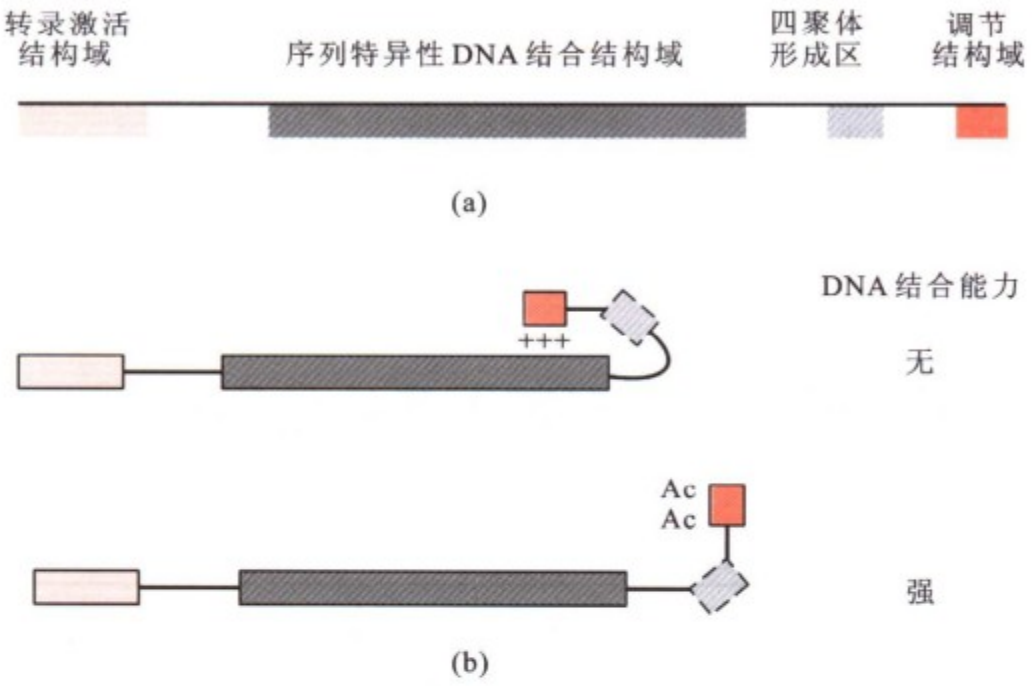



**图 8–40 p53 蛋白乙酰化有利于激活靶基因转录** (a)p53 蛋白基本结构示意图;(b)p53 蛋白的乙酰化提高了它的 DNA 结合能力。

视网膜母细胞瘤(retinoblastoma,Rb)蛋白是另一类被广泛研究的肿瘤抑制因子,该蛋白主要与 E2F 类转录激活因子相互作用,抑制靶基因(参与细胞周期调控)的转录活性。研究发现,与 E2F 类转录激活因子相结合的 Rb 能进一步募集去乙酰化酶 1(HDAC1),使这一类基因的启动子区发生特异性的去乙酰化反应,导致该区段染色质浓缩,靶基因转录活性消失(图 8–41)。有实验表明,DNA 甲基化诱导了组蛋白的去乙酰化作用,使靶基因转录受抑制(图 8–42)。

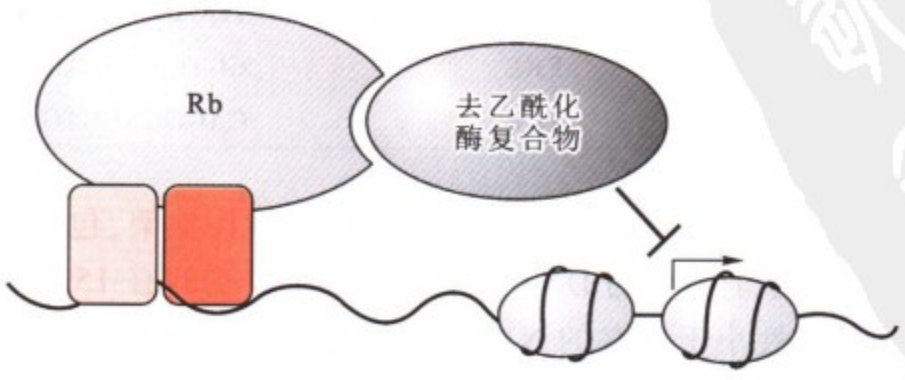



**图 8–41 组蛋白去乙酰化酶与肿瘤抑制因子 Rb 相互作用抑制靶基因转录**

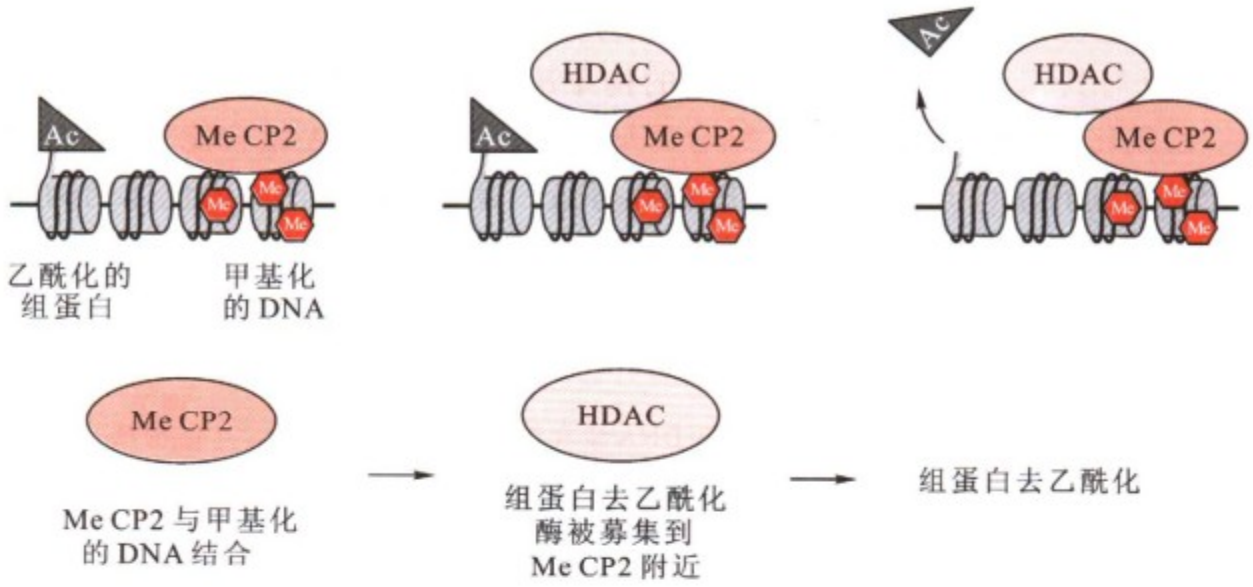



图 8-42 DNA 甲基化诱导组蛋白去乙酰化,从而抑制靶基因转录

## 8.5 激素与热激蛋白对基因表达的影响

### 8.5.1 激素对靶基因的影响

许多类固醇激素(如雌激素、孕激素、醛固酮、糖皮质激素和雄激素,图 8-43)以及一般代谢性激素(如胰岛素)的调控作用都是通过起始基因转录而实现的。靶细胞具有专一的细胞质受体,可与激素形成复合物,导致三维结构甚至化学性质的变化。经修饰的受体与激素复合物通过核膜进入细胞核内,并与染色质的特定区域结合,导致基因转录的起始或关闭。研究发现,体内存在的许多糖皮质类激素应答基因都有一段大约 20bp 的顺式作用元件(激素应答元件,简称 HRE),该序列具有类似增强子的作用,其活性受激素制约。表 8-13 是 3 种激素应答序列成分保守性分析。靶细胞中含有大量激素受体蛋白,而非靶细胞中没有或很少有这类受体,这是激素调节转录组织特异性的根本原因。

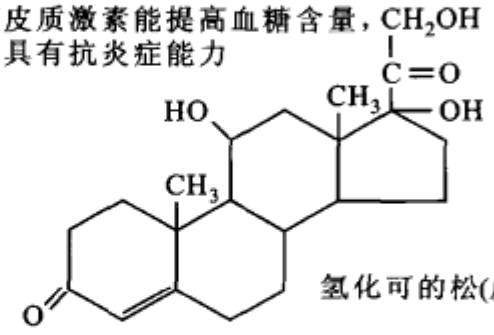
表 8-13 三种激素应答序列成分分析

| 激素 | 元件 | 序列 |
|---|---|---|
| 糖皮质激素 | GRE | TGGTACANNNTGTTCT |
| 雌激素 | ERE | GGTCANNNTGTCC |
| 甲状腺素 | TRE | CAGGGACGTGACCGCA |

所有固醇类激素的受体蛋白分子都有相同的结构框架,包括保守性极高的(42%~94%)、位于分子中央的 DNA 结合区,位于 C 端的有 15%~57%同源性的激素结合区和保守性小于 15%的 N 端。该区的具体功能不详,但它的存在保证了转录的高效进行。研究还发现,如果糖皮质激素受体蛋白激素结合区的某个部分丢失,就变成一种永久性的活性分子,即无需激素诱导也有激活基因转录的作用(图 8-44)。研究表明,激素、受体与顺式作用元件的结合位点三者缺一不

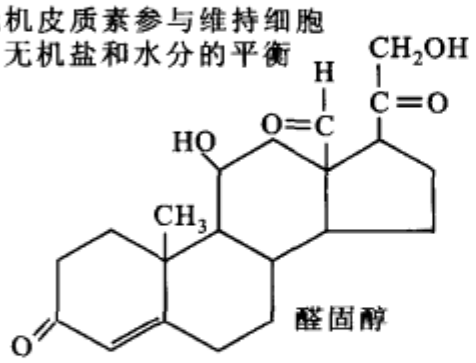
**类皮质激素(肾上腺皮质激素)**

糖皮质激素能提高血糖含量，并具有抗炎症能力




氢化可的松(皮质醇)

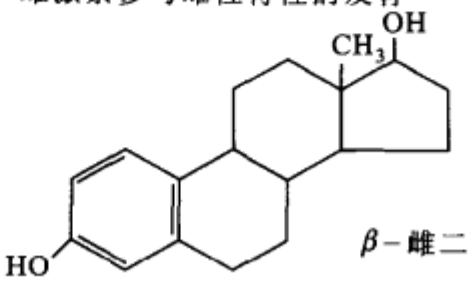
无机皮质素参与维持细胞内无机盐和水分的平衡




醛固酮

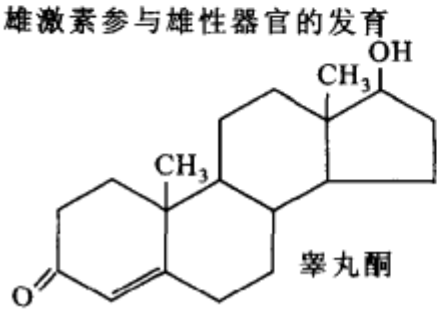
**固醇类性激素**

雌激素参与雌性特性的发育




β－雌二醇

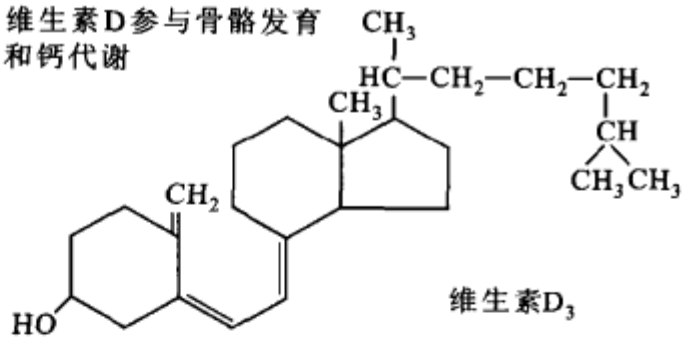
雄激素参与雄性器官的发育




睾丸酮

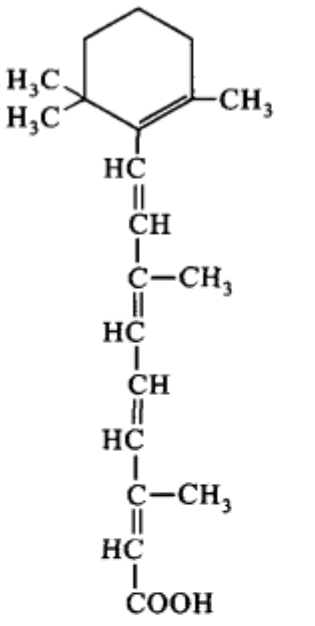
**其他**

维生素D参与骨骼发育和钙代谢




维生素$D_3$

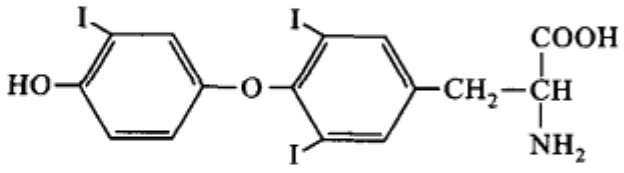
视黄酸参与形态建成




反式视黄酸

甲状腺激素

甲状腺激素调控了体内的基础代谢水平




三碘甲腺原氨酸

▌图 8–43 几种常见的疏水性小分子激素的结构式

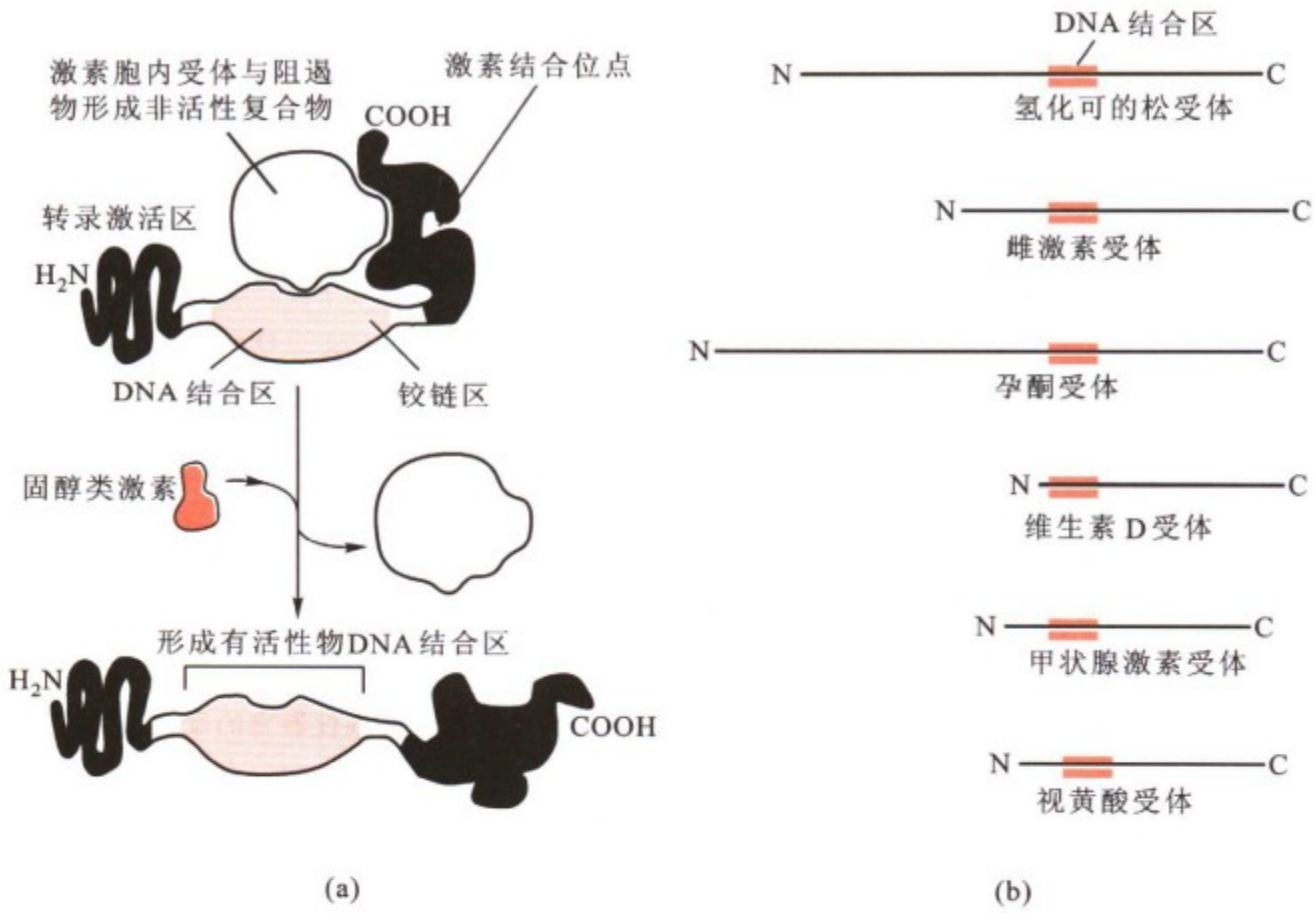



**图 8–44 激素及其胞内受体介导的基因表达调控模式** (a)小分子激素激活其胞内受体分子的机制;(b)胞内受体的基本结构分析。

可,其中无论是受体蛋白与激素的结合,还是激素本身,都不是与 DNA 结合并激活转录所必需的。其实,通常情况下,受体蛋白中激素结合结构域妨碍了DNA结合区及转录调控区发挥生理功能,只有与相应激素结合后才能打破这种障碍。

雌激素刺激鸡输卵管中卵清蛋白合成可能是激素诱导转录作用的典型例子。将雌激素注入鸡体内,输卵管组织便开始合成卵清蛋白 mRNA。只要有雌激素或类雌激素(如二乙基己烯雌酚)存在,合成便能继续下去。一旦停止供应激素,合成速度便下降直至完全消失。多肽激素胰高血糖素接触靶细胞时,首先与受体结合,并激活膜上的腺苷酸环化酶,使之以 $Mg^{2+}$–ATP 为底物生成环腺苷酸和焦磷酸。细胞内 cAMP 浓度升高,导致蛋白激酶活性增强,特定酶系的磷酸化水平及酶活性都得到改善,促进糖原最终分解为葡糖–1–磷酸(图 8–45)。

激素受体对各自特定的激素具有高度特异的识别能力及亲和力,激素到达靶细胞时能迅速与受体结合生成激素–受体复合物,诱导生理生物化学效应。有 3 种较为流行的假说:

(1) 受体流动假说 当激素与受体结合时,激素–受体复合物可在膜内移动并与腺苷酸环化酶结合,激活环化酶,引发后续效应。通过激素的稀释或解离,使受体和环化酶回复到非偶联状态。这一假说强调了激素–受体复合物在膜内的随机碰撞受环境因素,如温度、pH、离子浓度及膜脂中磷脂和胆固醇比例变化所制约。

(2) 中介物假说 研究腺苷酸环化酶和胰高血糖素受体发现,受体与酶之

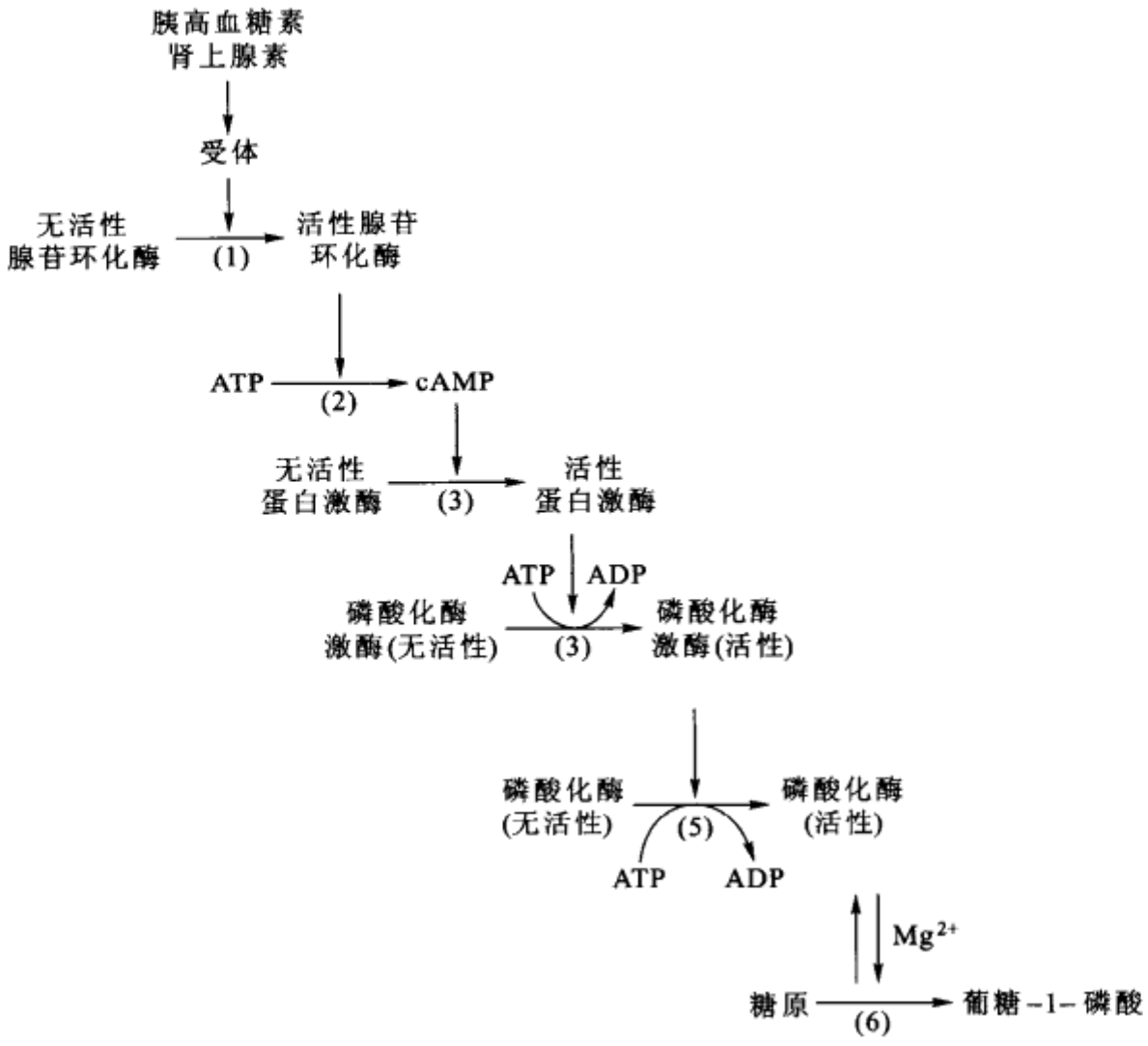



图 8-45 胰高血糖素激活糖原分解连锁反应

间可能通过膜脂偶联在一起。用适量的高纯磷脂酶 A 处理肝细胞膜，去脂后，氟离子仍可刺激膜上的腺苷酸环化酶活性，但对胰高血糖素的作用则完全丧失。若加入膜脂则激素的作用恢复，说明去膜脂后激素的信息传递中断，而对氟离子的作用不受影响。进一步研究发现，所加入的磷脂具有特异性。如加入磷脂酰丝氨酸，仅可恢复胰高血糖激素激活酶的作用，而对肾上腺素无作用。只有加磷脂酰肌醇，才能恢复肾上腺素激活腺苷酸环化酶的效应。因此，膜上磷脂成分对这两个分子的偶联有着重要的作用。

(3) 邻位互调假说　有科学家根据 GTP 能影响激素与受体结合并进而影响腺苷酸环化酶活性的现象，认为该酶系统至少存在 3 个活性部位，即激素-受体结合部位、$Mg^{2+}$-ATP 作用中心和核苷酸调节部位。腺苷酸环化酶有几种不同的立体构象，激素与 GTP 协同作用可使这几种不同的构象之间发生互变。

## 8.5.2 热休克蛋白诱导的基因表达

现代分子生物学上把能与某个(类)专一蛋白因子结合，从而控制基因特异表达的 DNA 上游序列称为应答元件(response element)，如热休克应答元件(heat shock response element，HSE)，糖皮质应答元件(glucocorticoid response element，GRE)，金属应答元件(metal response element，MRE)等，这些应答元件与细胞内

专一的转录因子相互作用,协调相关基因的转录(表 8-14)。

表 8-14 特异转录因子与各类应答元件的相互作用

| 调控因子 | 应答元件 | DNA 序列 | 结合蛋白 | 相对分子质量 |
|---|---|---|---|---|
| 热休克 | HSE | CNNGAANNTCCNNG | HSF | $9.3\times10^4$ |
| 镉 | MRE | CGNCCCGGNCNC | ? | ? |
| 佛波酯 | TRE | TGACTCA | API | $3.9\times10^4$ |
| 血清 | SRE | CCATATTAGG | SRF | $5.2\times10^4$ |

许多生物在最适温度范围以上，能受热诱导合成一系列热休克蛋白(heat shock protein,HSP),又称热激蛋白。受热后,果蝇细胞内 *hsp70* mRNA 水平提高 1 000 倍,就是因为热激因子(heat shock factor,HSF)与 *hsp70* 基因 TATA 区上游 60 bp 处的 HSE 结合,诱发转录起始。

按相对分子质量的大小以及同源程度可将热休克蛋白分为 HSP90、HSP70、小分子 HSP 及泛蛋白 4 个家族，各家族 HSP 又由多种不同形式或经不同修饰的蛋白质分子所组成。HSP 分布广泛,诱导迅速,即使在非热休克细胞中也存在着由热休克同源基因编码的类似蛋白质。真核细胞的热休克蛋白可能具有机体保护功能并在细胞的正常生长和发育中起重要作用,HSP70 的主要功能是参与蛋白质的代谢,而泛蛋白的主要功能是清除细胞内的变性蛋白质。HSP 还常与具有不同功能的多种蛋白质形成天然复合物,参与有关蛋白质折叠、亚基的组装、细胞内运输以及降解等过程。热休克蛋白参与靶蛋白活性和功能的调节,却不是靶蛋白的组成部分,因此,一般称它为分子伴侣或伴侣蛋白(chaperonin)。现将主要 HSP 家族及其生理功能列于表 8-15。

表 8-15 主要 HSP 家族及其生理功能

| 家族 | 主要成员 | 主要生理功能 | 免疫应答中的可能作用 |
|---|---|---|---|
| HSP90 | HSP90,HSP83 | 促进甾醇激素受体与激素的结合及与 DNA 结合,调节激酶磷酸化活性 | 抗肿瘤,提高自身免疫功能 |
| HSP70 | HSP70, BiP, Dnat, Hsc70,Grp78 | 参与蛋白质折叠和去折叠、蛋白转位及多聚复合物的组装 | 免疫球蛋白装配,类抗原加工,病原体抗原及自身免疫性诱导 |
| 小分子 HSP | Hsp27,Gro23,HSP16 | 参与蛋白质折叠和去折叠及多聚复合物的装配 | 病原体抗原及自身免疫性诱导 |
| 遍在蛋白质 | 遍在蛋白质 | 蛋白质降解 | 类抗原加工,淋巴细胞回巢,自身免疫性诱导 |

*hsp* 基因中内含子数量都很少,至今为止尚未发现 *hsp70* 基因中有内含子,而人 hsp90、果蝇 hsp82、人 hsp27、鸡 hsp108 和泛蛋白等都只有少量内含子。*hsp* 的这一基本特征保证了它们一旦起始转录，不需剪接就可产生出成熟

mRNA 以适应 *hsp* 大量快速表达的需要，防止严重的热休克影响 mRNA 前体的剪接。

在没有受热或其他环境胁迫时，HSF 主要以单体的形式存在于细胞质和核内。单体 HSF 没有 DNA 结合能力，HSP70 可能参与了维持 HSF 的单体形式。受到热激或其他环境胁迫时，细胞内变性蛋白增多，它们都与 HSF 竞争结合 HSP70，从而释放 HSF，使之形成三体并输入核内。HSF 一旦形成三体，便拥有与 HSE 特异结合、促进基因转录的功能。这种能力可能还受磷酸化水平的影响，因为热激以后，HSF 不但形成三体，还会迅速被磷酸化。HSF 与 HSE 的特异性结合，引起包括 HSP70 在内的许多热休克应答基因表达，大量产生 HSP70 蛋白。随着热激温度消失，细胞内出现大量游离的 HSP70 蛋白，它们与 HSF 相结合，形成没有 DNA 结合能力的单体并脱离 DNA(图 8-46)。

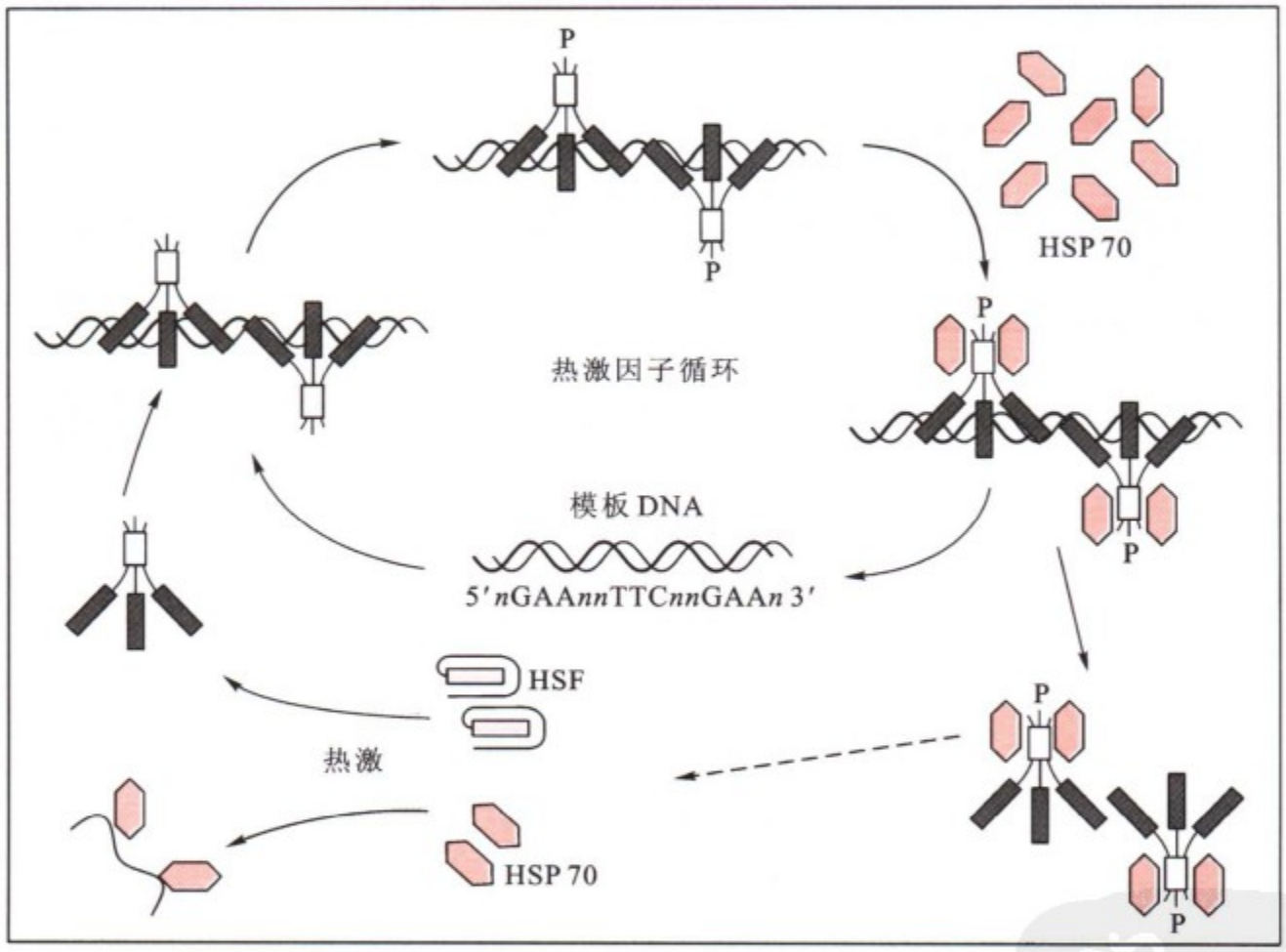



图 8-46 热休克蛋白调控的基因表达机制

分析果蝇和人 HSF 蛋白表明，热休克转录因子具有多个可形成拉链的疏水 DNA 区域，其中 3 个位于 N 端，靠近 DNA 结合区，参与促进 HSF 三体的形成。第 4 个拉链位于 C 端，与位于第 452~488 位氨基酸残基的保守区，都是维持 HSF 单体构象所必需的。序列删除实验表明，无论去除第 4 个拉链区，还是更换该区甲硫氨酸 $^{391}$→赖氨酸，亮氨酸 $^{395}$→脯氨酸，都会导致 HSF 突变体对 HSE 在常温下的高亲和力。删除第 452~488 位氨基酸，也可部分替代热休克效应(图 8-47)。根据以上实验结果，有研究者提出如图 8-48 的模式，认为在常温下，HSP 蛋白第 1 和 4 号拉链发生相互作用，从而阻断了三体的形成；热激或其他

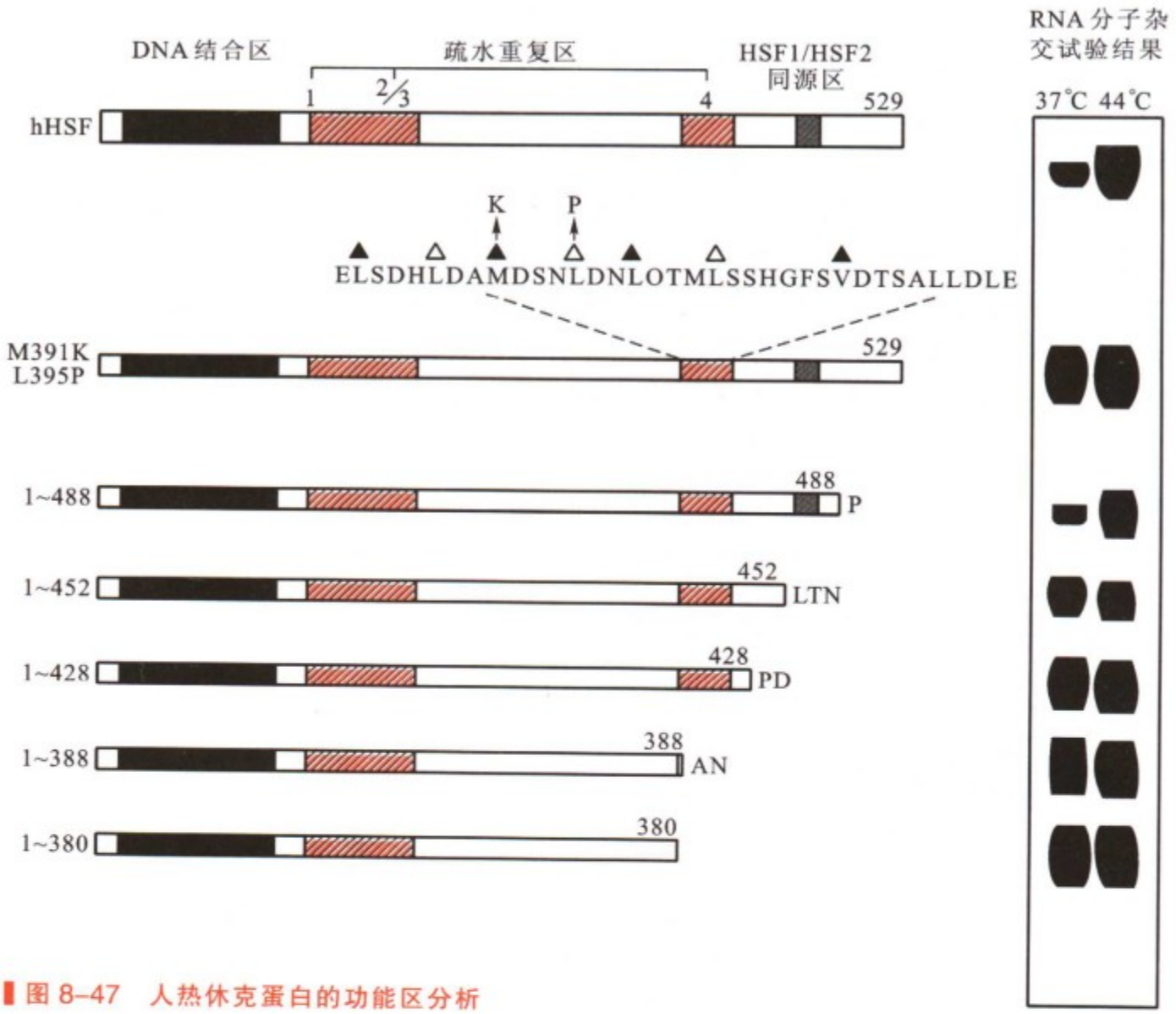



图 8-47 人热休克蛋白的功能区分析

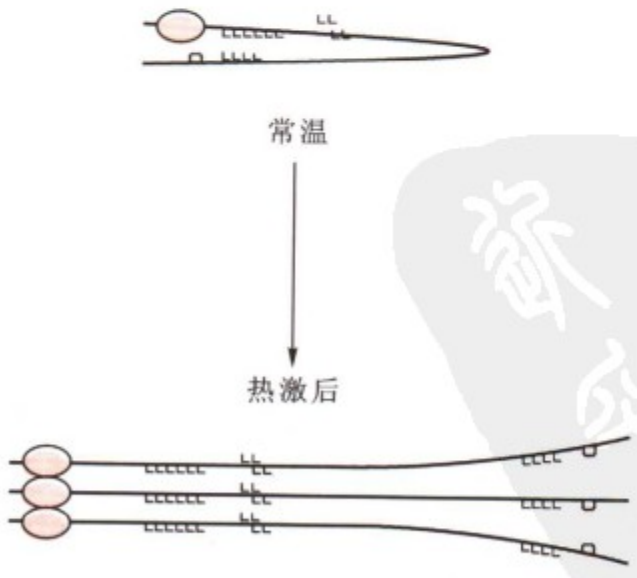



图 8-48 HSF 在常温及热激后的构象变化

环境胁迫导致内源拉链之间的解离,蛋白质伸展成长链,不同链上的拉链发生作用,形成具有 DNA 结合能力的三体。

## 8.6 其他水平上的表达调控

尽管转录调节可能是基因表达调控的最重要方式,但并不是唯一的方式。因为真核生物基因转录在核内进行,而翻译(蛋白质合成)却在细胞质中进行,加上真核基因有插入序列,结构基因被分割成不同的片段,基因的转录后调控就显得更为重要。除了前面介绍的 RNA 剪接外,RNA 的加工成熟和蛋白质合成,都在真核基因调控中起着不可忽视的作用。

### 8.6.1 RNA 的加工成熟

各种基因的转录产物都是 RNA,无论是 rRNA、tRNA 还是 mRNA,初级转录产物只有经过加工,才能成为有生物功能的活性分子。

1. rRNA 和 tRNA 的加工成熟

rRNA 加工有两个内容,一个是分子内的切割,另一个是化学修饰。真核生物的 *rRNA* 基因转录时先产生一个 45S 的前体 rRNA,然后被核酸酶逐渐降解,形成成熟的 18S、28S 和 5.8S rRNA。

*rRNA* 基因转录主要在细胞核仁内进行,初级转录产物 45S 前体 rRNA 很快就会被加工降解,生成不同相对分子质量的成熟 rRNA。rRNA 的化学修饰主要是甲基化,原核生物 rRNA 主要是碱基甲基化,而真核生物 rRNA 则主要是核糖甲基化。*tRNA* 基因转录时也可能先生成前体 tRNA,然后再进行加工成熟。一般认为,*tRNA* 基因的初级转录产物在进入细胞质后,首先经过核苷的修饰,生成 4.5S 前体 tRNA,再行剪接成为成熟 tRNA(4S)。

2. mRNA 的加工成熟

编码蛋白质的基因转录产生 mRNA。这类基因产物在转录后要进行一系列的加工变化,才能成为成熟的有生物功能的 mRNA。这些加工主要包括在 mRNA 的 5′末端加“帽子”,在其 3′端加上 poly A,进行 RNA 的剪接以及核苷酸的甲基化修饰等。由于 mRNA 的这些结构与它作为蛋白质合成模板的功能有密切关系,所以是基因表达的重要调控环节。

与 rRNA、tRNA 相同,编码蛋白质的基因转录时首先生成 pre-mRNA(或称核不均一 RNA,hnRNA),然后再加工剪接为成熟 mRNA。如下几个方面的证据说明 hnRNA 确是 mRNA 的前体。

① 在真核生物细胞核中发现存在代谢十分活跃、平均相对分子质量为 $2\times10^7$,长度不均一的 RNA,而细胞质中 mRNA 的平均相对分子质量仅为 $1.5\times10^6$。β-珠蛋白基因的初级转录产物为 14S,而成熟珠蛋白 mRNA 只有 9S,说明在 mRNA 的加工成熟过程中有相当一部分核苷酸被删除。

② hnRNA 很不稳定,半衰期只有 5~15 min,而真核生物 mRNA 的半衰期一般都比较长,这说明 hnRNA 不可能是转录的最终产物,而只能是中间产物。

③ 用放射性同位素作脉冲标记证明,75%被标记的 RNA 是 hnRNA。这些标记的 RNA 大部分位于核内,只有 10%进入细胞质,所以认为它们是 mRNA 的前体。

④ hnRNA 占细胞全部 RNA 的 3%,说明它一经诞生就立即加工变化,不会长久积累下来。

⑤ hnRNA 3′端也带有 poly A。用 3H-AR(放射性同位素氚标记的腺嘌呤核苷)作脉冲标记实验,45 s 后发现,90%的 poly A RNA 位于细胞核内;标记 150 min 后,有 20%的 poly A RNA 进入细胞质,说明带 poly A 的 RNA 是由细胞核转运入细胞质的。

⑥ 加入 dAR(脱氧腺嘌呤核苷)抑制 poly A 的生物合成,hnRNA 和 mRNA 的合成都受到抑制。据此,科学家认为 hnRNA 是 mRNA 的前体,成熟 mRNA 是由 hnRNA 经剪接加工得到的。

3. 真核生物基因转录后加工的多样性

真核生物的基因可以按其转录方式分为两大类:即简单转录单位和复杂转录单位。这两种转录方式虽然最终都产生蛋白质,但它们的转录后加工方式是不同的。

(1) 简单转录单位　这类基因只编码产生一个多肽,其原始转录产物有时需要加工,有时则不需要加工。这类基因转录后加工有 3 种不同形式(图 8-49)。

第一种简单转录单位,如组蛋白基因,它们没有内含子,因此不存在转录后加工问题,其 mRNA 3′端没有 poly A,但有一个保守的回文序列作为转录终止信号。已知高等真核生物有 3 种不同的组蛋白,即与 DNA 复制有关的、与复制无关的和组织特异性组蛋白。在细胞分裂的 S 期(DNA 合成期)表达的、与复制

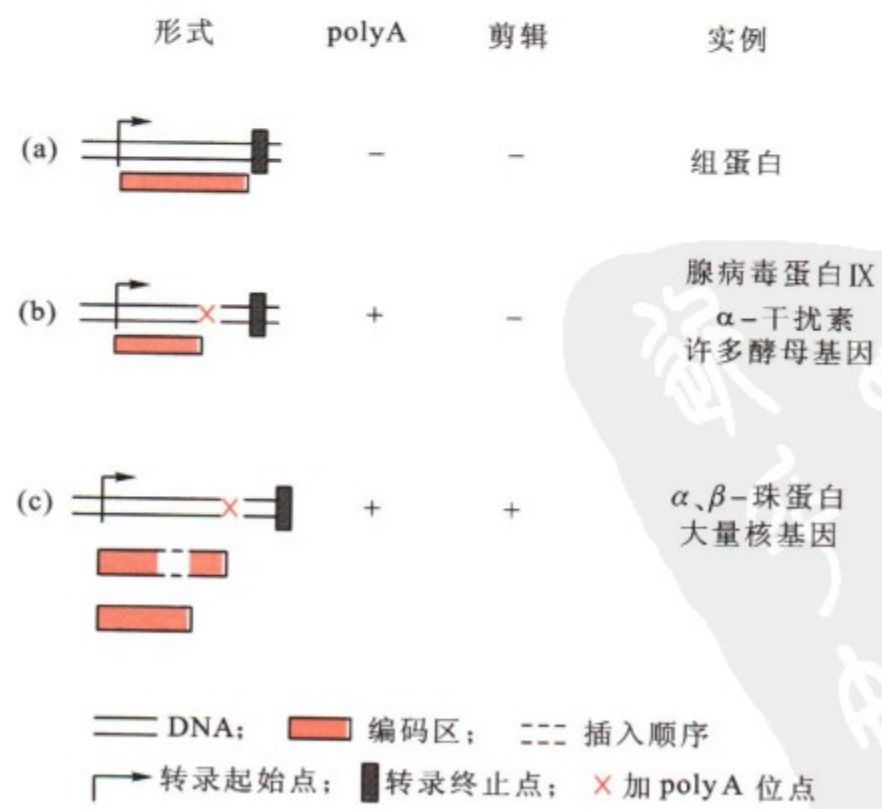



图 8-49　真核生物基因转录后加工的 3 种主要形式

有关的组蛋白是一个大家族，真核细胞中共有 5~20 个各自成簇的 $H_2A$、$H_2B$、$H_3$、$H_4$ 和 $H_1$ 基因。与复制无关的组蛋白(主要存在于不再分裂、已经完成分化的组织中)和组织特异性组蛋白(如鸟类和两栖动物红细胞专一的组蛋白 $H_5$)是由单拷贝基因编码的，这些单拷贝基因与其他组蛋白基因无关，由这些基因转录的 mRNA 是带有 poly A 尾巴的。

第二种简单转录单位包括腺病毒蛋白Ⅸ，α-干扰素和许多酵母蛋白质基因，它们没有内含子，所编码的 mRNA 不需要剪接，但需要加 poly A。

第三种简单转录单位包括 α-和 β-珠蛋白基因及许多细胞蛋白基因，这些基因虽然都有内含子，需要进行转录后加工剪接，还要加 poly A，但它们只产生一个有功能的 mRNA，所以仍然是简单转录单位。

(2) 复杂转录单位　含有复杂转录单位的主要是一些编码组织和发育特异性蛋白质的基因，它们除了含有数量不等的内含子以外，其原始转录产物能通过多种不同方式加工成两个或两个以上的 mRNA。

① 利用多个 5′端转录起始位点或剪接位点产生不同的蛋白质。小鼠和鸡的肌球蛋白轻链基因、果蝇 *EH8* 基因、脱氢酶基因、小鼠唾液与肝脏 α-淀粉酶基因、田鼠乙酰辅酶 A 基因等都属于这种成熟方式。图 8-50 说明肌球蛋白碱性轻链基因选用了不同的 5′转录起始位点及剪接不同外显子产生蛋白质异构体 LC1 和 LC3 的过程。由于选用不同的转录起始位点，还导致了小鼠 α-淀粉酶基因的组织特异性。

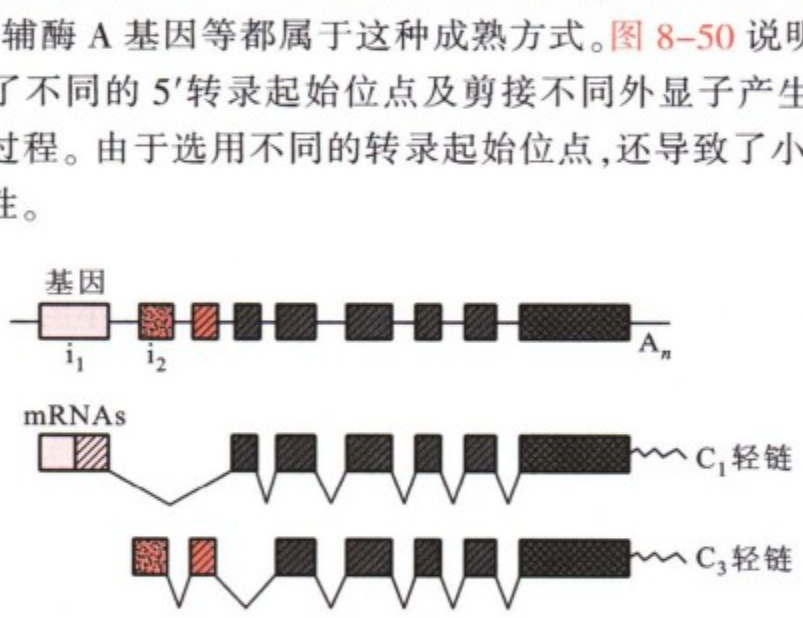



图 8-50　肌球蛋白碱性轻链基因剪接的多样性

② 利用多个加 poly A 位点和不同的剪接方式产生不同的蛋白质。这类基因 5′端虽只有一个转录起始位点，但有两个或多个加 poly A 位点，因此可通过不同的剪接方式得到不同的蛋白质。属于这类基因的有腺病毒晚期基因，高等真核生物免疫球蛋白 M、D、E 和 G 的 H 链基因，大鼠降钙素基因，牛前激肽原基因，果蝇肌球蛋白 H 链基因，果蝇肌球蛋白碱性 L 链基因和人血纤维蛋白原基因等。以大鼠降钙素基因为例，它编码与血钙代谢有关的蛋白质。该基因有一个转录起始位点和 6 个外显子，两个加 poly A 位点。如果初级转录产物以 0—1—2—3—4 号外显子连接在一起，并利用外显子 4 后面的 poly A 位点，就产生降钙素 mRNA。但当初级转录产物以 0—1—2—3—5—6 号外显子连接在一起，并利用外显子 6 以后的加 poly A 位点时，就产生了编码与降钙素基因表达有关的多肽 CGRP 的 mRNA(图 8-51)。

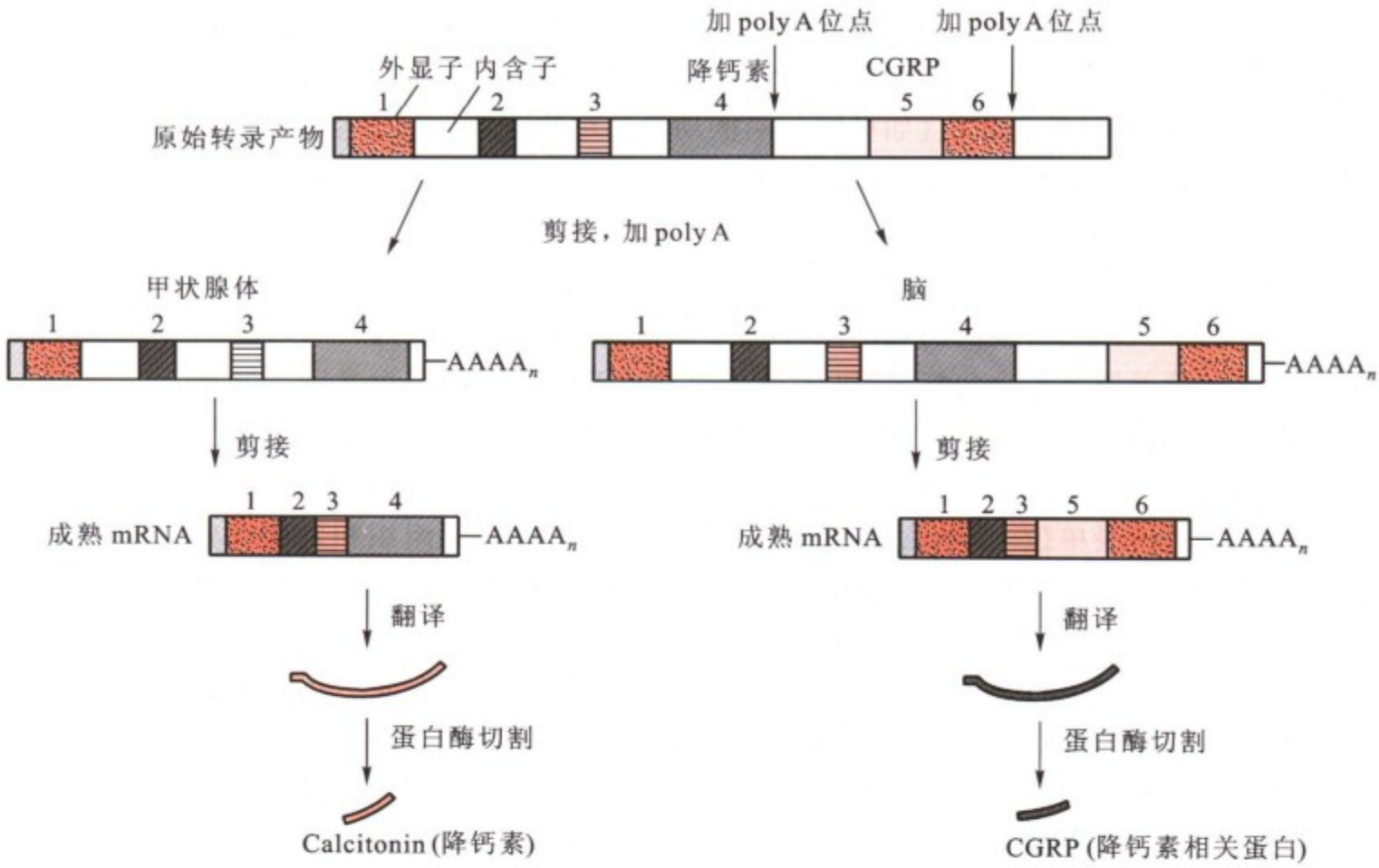



图 8-51 pre-mRNA 不同的剪接方式造成了不同组织中不同的降钙素样蛋白

免疫球蛋白的 μ 链基因也有一个转录起始点和两个 poly A 位点。当它启用外显子 6 上的终止序列时，翻译生成相对分子质量较小的分泌型蛋白(μs)，可直接分泌进入细胞质中；当启用外显子 8 上的终止序列时，翻译生成相对分子质量较大的膜结合型蛋白(μm)，它能与细胞膜结合在一起。

③ 虽无剪接，但有多个转录起始位点或加 poly A 位点的基因，如二氢叶酸还原酶基因和酵母乙醇脱氢酶基因都具有不同的 5′端和 poly A 位点。大鼠 α-2μ-珠蛋白基因、鸡波形蛋白基因、*X* 基因和人 *N-ras* 基因均含有多个 poly A 位点，而鸡溶菌酶基因和酵母蔗糖酶基因则拥有多个 5′端。这类基因表达时，往往产生不同 5′端的 mRNA。虽然原始转录产物的蛋白质编码区和 3′端都相同，不同的 5′端却导致了分泌型和胞内型蛋白的产生，其中分泌型蔗糖酶是可调节的，而胞内型则是组成型的。

4. mRNA 有效性的调控

真核生物能否长时间、及时地利用成熟的 mRNA 分子翻译出蛋白质以供生长、发育的需要，是与 mRNA 的稳定性以及屏蔽状态的解除密切相关的。原核生物 mRNA 的半衰期很短，平均大约 3 min。高等真核生物迅速生长的细胞中 mRNA 的半衰期平均约为 3 h。在高度分化的终端细胞中许多 mRNA 极其稳定，有的寿命长达几天或十几天，加上强启动子的转录，使一些终端细胞特有的蛋白质合成达到惊人的水平。例如，家蚕丝心蛋白基因具有很强的启动子，几天内即可转录出 $10^5$ 个丝心蛋白 mRNA，而它的寿命长达 4 d，每个 mRNA 分子能重复翻译出 $10^5$ 个丝心蛋白，所以 4 d 内可产生 $10^{10}$ 个丝心蛋白，说明 mRNA 寿命

的延长是 mRNA 有效性的一个重要因素。

### 8.6.2 翻译水平的调控

在蛋白质生物合成的起始反应中主要涉及细胞中的 4 种装置,包括:①核糖体　它是蛋白质生物合成的场所;②蛋白质合成的模板——mRNA　它是传递基因信息的媒介;③可溶性蛋白因子　这是蛋白质生物合成起始物形成所必需的因子;④tRNA　它是氨基酸的携带者。只有这些装置和谐统一才能完成蛋白质的生物合成。

1. 真核生物 mRNA 的“扫描模式”与蛋白质合成的起始

大量实验证明,在真核生物起始蛋白质合成时,40S 核糖体亚基及有关合成起始因子首先与 mRNA 模板靠近 5′端处结合,然后向 3′方向滑行,发现 AUG 起始密码子时,与 60S 大亚基结合形成 80S 起始复合物。这就是 Kozak 提出的真核生物蛋白质合成起始的“扫描模式”。

那么,为什么核糖体滑行到 mRNA 的第一个 AUG,即在离 5′端最近的起始密码位点就停下来并起始翻译呢?现在认为,这是由 AUG 的前(5′方向)和后(3′方向)序列所决定的。调查了 200 多种真核生物 mRNA 中 5′端第一个 AUG 前后序列发现,除少数例外,绝大部分都是 A/G NNAUGG,说明这样的序列对翻译起始来说是最为合适的。所以,“扫描模式”相当合理地说明了许多真核生物 mRNA 的单一顺反子性质,也合理地解释了为什么将 mRNA 水解之后,它内部的起始密码会被活化。因为 mRNA 的水解产生出了新的 5′端,所以它又可以成为 40S 起始复合物的进入位点。

那么,真核生物中具有相关生物功能的基因通过什么样的补偿机制得以协同表达呢?有人提出,真核生物通过“融合基因”的方式产生出多聚顺反子的替代物,即“多聚蛋白质”。多聚蛋白质是由相当长的 mRNA 编码的,如色氨酸合成酶,它的 α 亚基(相对分子质量 28 727)和 $\beta_2$ 亚基(相对分子质量 42 756)在大肠杆菌中分别由不同基因编码,但是,在真核生物酵母中,这两个多肽融合形成相对分子质量为 76 000 的双功能蛋白质。红色链孢霉的 *arom* 基因簇也编码多功能蛋白,参与多聚芳香化合物的生物合成。此外,酵母的 *his4* 基因编码了一个三功能蛋白质。从功能上说,哺乳动物的脂肪酸合成酶(相对分子质量 $2.4\times10^5$)相当于大肠杆菌中一组 7 个之多的功能各不相同的多肽。所以,基因融合可能补偿了真核生物核糖体不能利用多聚顺反子型 mRNA 来协调各种不同蛋白质生物合成的缺陷。

2. mRNA 5′端帽子结构的识别与蛋白质合成

因为绝大多数真核生物 mRNA 5′端都带有“帽子”结构,所以,核糖体起始蛋白质的合成,首先面临的问题是如何识别这顶“帽子”。

真核生物 mRNA 5′端可有 3 种不同的帽子,即 O 型、Ⅰ型和Ⅱ型,其主要差异在于帽子中碱基甲基化程度的不同。表 8-16 列举了这 3 种不同类型帽子的结构。

真核生物的加帽子反应发生在 mRNA 前体刚转录出来不久或尚未转录完

表 8-16 真核生物 mRNA 的帽子结构

| 种类 | 结构 | mRNA |
|---|---|---|
| 0 型 | $m^7GpppA/GpNp-$ | 酵母、黏菌 |
| Ⅰ型 | $m^7GpppA_{mp}Gm\ pNp-$ | 海胆胚、卤虫 |
| Ⅱ型 | $m^7GpppN_{1mp}N_{2p}$ | 哺乳动物 |
| | $m^7GpppN_{1mp}N_{2mp}$ | 哺乳动物 |

成时,催化这一过程的是鸟苷酸转移酶和甲基转移酶,它们都位于细胞核内。据认为,细胞中的帽子化酶是同时具有 RNA 三磷酸酯酶和鸟苷酸转移酶活性的多功能酶。通过鸟苷酸转移酶生成的是 5′-5′磷酸二酯键。从 0 型到Ⅰ型帽子的生成都在细胞核内进行,由Ⅰ型帽子进一步加工成Ⅱ型帽子在细胞质内进行。

3. mRNA 的稳定性与基因表达调控

在高等真核生物中转铁蛋白受体(TfR)和铁蛋白负责铁吸收和铁解毒。这两个 mRNA 上存在相似的顺式作用元件,称为铁应答元件(iron responsive element,IRE),IRE 与 IRE 结合蛋白(IRP)相互作用控制了这两个 mRNA 的翻译效率。当细胞处于缺铁或高铁水平时,能产生两个数量级的蛋白水平差异,却没有在 mRNA 水平上发现存在显著差异。研究表明,位于 5′非翻译区的 IRE 控制了铁蛋白 mRNA 的翻译效率,去掉这个非翻译区 IRE,可造成铁蛋白的永久性高水平翻译。当细胞缺铁时,IRP 与 IRE 具有高亲和力,两者的结合有效地阻止了铁蛋白 mRNA 的翻译,与此同时,TfR mRNA 上 3′非翻译区中的 IRE 也与 IRP 特异结合,有效地阻止 TfR mRNA 的降解,促进 TfR 蛋白的合成(图 8-52)。此外,催乳素能明显延缓酪蛋白(casein)mRNA 的降解。不加入催乳素时,体系中的酪蛋白 mRNA 在 1 h 内降解 50%,而加入催乳素后 40 h,酪蛋白 mRNA 才降解 50%(图 8-53)。

4. 可溶性蛋白因子的修饰与翻译起始调控

许多可溶性蛋白因子,即起始因子,对蛋白质合成的起始有着重要的作用,对这些因子的修饰也会影响翻译起始。

(1) eIF-2 磷酸化对翻译起始的影响 用兔网织红细胞粗抽提液研究蛋白质合成时发现,如果不向这一体系中添加氯高铁血红素,几分钟之内蛋白质合成活性急剧下降,直到完全消失。这就是说,当没有氯高铁血红素存在时,网织红细胞粗抽提液中的蛋白质合成抑制剂就被活化,从而抑制蛋白质合成。现已查明,该抑制剂 HCI 是受氯高铁血红素调节的,它其实是 eIF-2 的激酶,可以使eIF-2 的 α 亚基磷酸化,并由活性型变成非活性型。没有生物活性的 HCI 也可以通过自身的磷酸化变成活性型,这个过程可能是自我催化的,并与一个被称为 HS 因子的热稳定蛋白有关。氯高铁血红素能够阻断 HCI 的活化过程(图 8-54)。

实验表明,eIF-2 的 α 亚基磷酸化以后,eIF-2 与 eIF-2B 紧密结合,直接影响了 eIF-2 的再利用,从而抑制翻译,影响蛋白质合成起始复合物的生成。eIF-2 磷酸化阻碍蛋白质生物合成这一现象,同样可由添加氧化型谷胱甘肽或双链RNA 等引发。

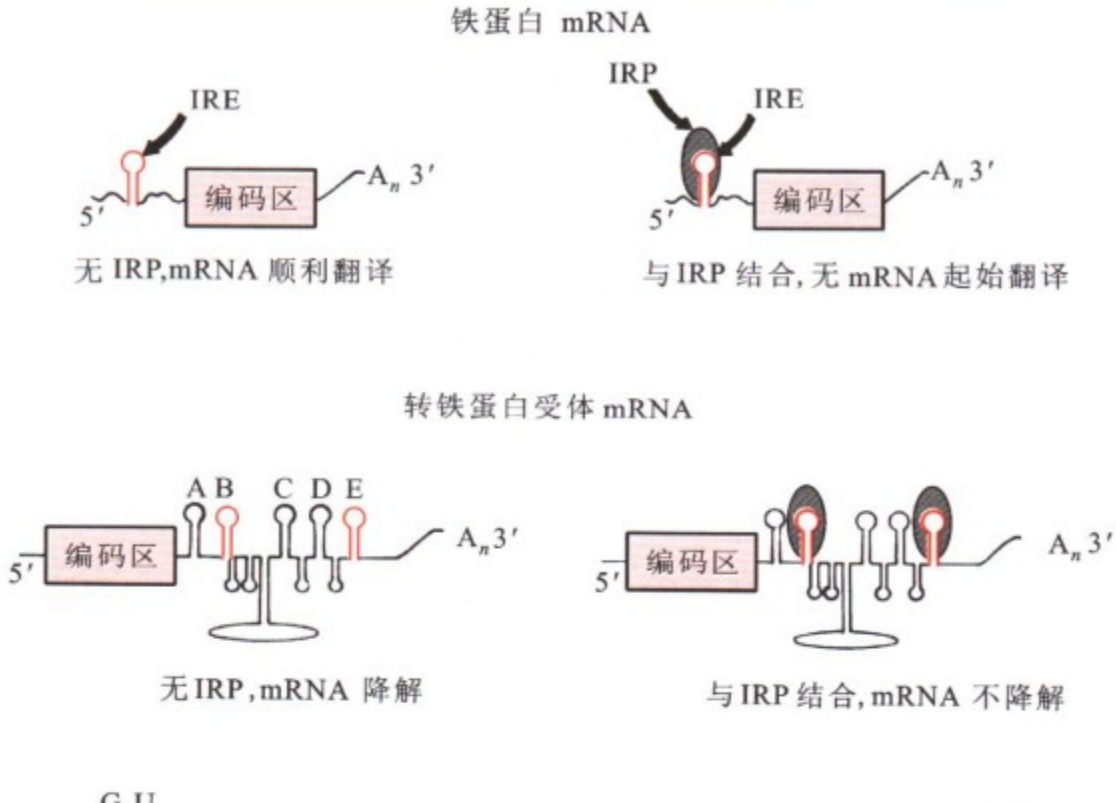



图 8-52 人铁蛋白及转铁蛋白受体 mRNA 的翻译调控机制

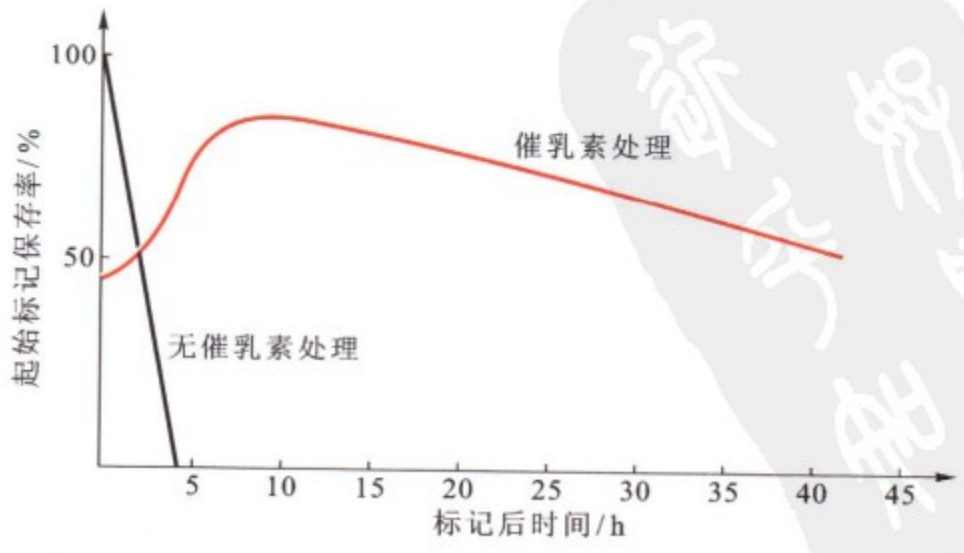
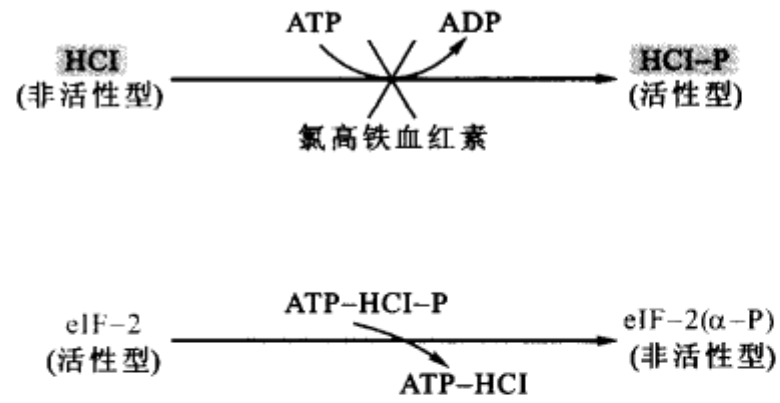



图 8-53 催乳素处理能明显延缓酪蛋白 mRNA 的降解




▌图 8-54 eIF-2、HCI 及氯高铁血红素的相互关系

(2) CBPⅡ活性与翻译的起始 在脊髓灰质病毒感染的 HeLa 细胞中,有帽子结构的宿主 mRNA 的翻译受阻,宿主蛋白质合成停止,但没有帽子结构的脊髓灰质炎病毒 mRNA 的翻译却不受影响。研究表明,宿主细胞 CBPⅡ失活是导致这种 mRNA 选择性翻译的根本原因。如果在这种感染细胞抽提液的蛋白质合成体系中加入由兔网织红细胞中提取的 CBPⅡ,就能恢复有帽子 mRNA 的翻译活性。添加 CBPⅡ对脊髓灰质炎病毒 RNA 的翻译没有影响。

在蛋白质生物合成的过程中,特别是在起始反应中,mRNA 的"可翻译性"是起决定作用的,其 5′端的帽子结构、二级结构、与 rRNA 的互补性以及起始密码附近的核苷酸序列都是蛋白质生物合成的信号系统。蛋白质生物合成的调控,就是通过 mRNA 本身所固有的信号与可溶性蛋白因子或者与核糖体之间的相互作用而实现的。

## 思考题

1. 基因家族的分类及其主要表达调控模式。
2. 何为外显子,内含子及其结构特点和可变调控?
3. DNA 甲基化对基因表达的调控机制。
4. 真核生物转录元件组成及其分类。
5. 增强子的作用机制。
6. 反式作用因子的结构特点及其对基因表达的调控?
7. 举例说明蛋白质磷酸化如何影响基因表达。
8. 组蛋白乙酰化和去乙酰化影响基因转录的机制。
9. 激素影响基因表的基本模式。
10. 分子伴侣的分类及其影响基因表达的机制。

## 主要参考文献

1. Mohr I. To replicate or not to replicate: achieving selective oncolytic virus repli-

cation in cancer cells through translational control. Oncogene, 2005, 24: 7697–7709.

2. Struhl1 K. Histone acetylation and transcriptional regulatory mechanisms. Genes Development, 1998, 12: 599–606.

3. Sakaguchi K, *et al*. DNA damage activates p53 through a phosphorylation–acetylation cascade. Genes Development, 1998, 12: 2831–2841.

# 9 疾病与人类健康


- 9.1 肿瘤与癌症
- 9.2 人类免疫缺陷病毒——HIV
- 9.3 乙型肝炎病毒——HBV
- 9.4 人禽流感的分子生物学机制
- 9.5 严重急性呼吸系统综合征-SARS 的分子机制
- 9.6 基因治疗


## 9.1 肿瘤与癌症

癌(cancer)是一群不受生长调控而繁殖的细胞,也称恶性肿瘤。与此相对应的良性肿瘤是一群仅局限在自己的正常位置,且不侵犯周围其他组织和器官的细胞。因此,绝大多数癌是由肿瘤细胞经过一系列突变转化而来的(图 9-1)。到目前为止,癌症仍然是人类生存的头号敌人。今天,科学家已经能够对许多种癌症做出早期诊断,延长患者的生存期,但并未从根本上征服这个病魔。20 世纪 50 年代以来,对肿瘤医学的研究有三项最为突出的成就:①癌基因的发现及其研究的深入;②染色体畸变与致癌基因表达相互关系的揭示;③抑癌基因的发现及表达调控。已经发现了上百个原癌基因(proto-oncogene)和许多抑癌基因,证明细胞癌变的分子基础是基因突变,DNA 的变化和不正常活动导致了细胞癌变。

癌基因(oncogene)可分为两大类:一类是病毒癌基因(viral oncogene,V-onc),编码病毒癌基因的主要有 DNA 病毒和 RNA 病毒。DNA 病毒包括乙型肝炎病毒、SV40 和多瘤病毒、乳头瘤病毒、腺病毒、疱疹病毒和痘病毒。RNA 病毒主要是反转录病毒。反转录病毒致癌基因(retrovirus onc)可能是研究得最多的病毒基因,它们能使靶细胞发生恶性转化。第二类是细胞转化基因(C-onc),它们能使正常细胞转化为肿瘤细胞,这类基因与病毒癌基因有显著的序列相似性。事实上,细胞转化基因可能就是存在于人体正常细胞中的原癌基因的突变产物,它们广泛存在于生物

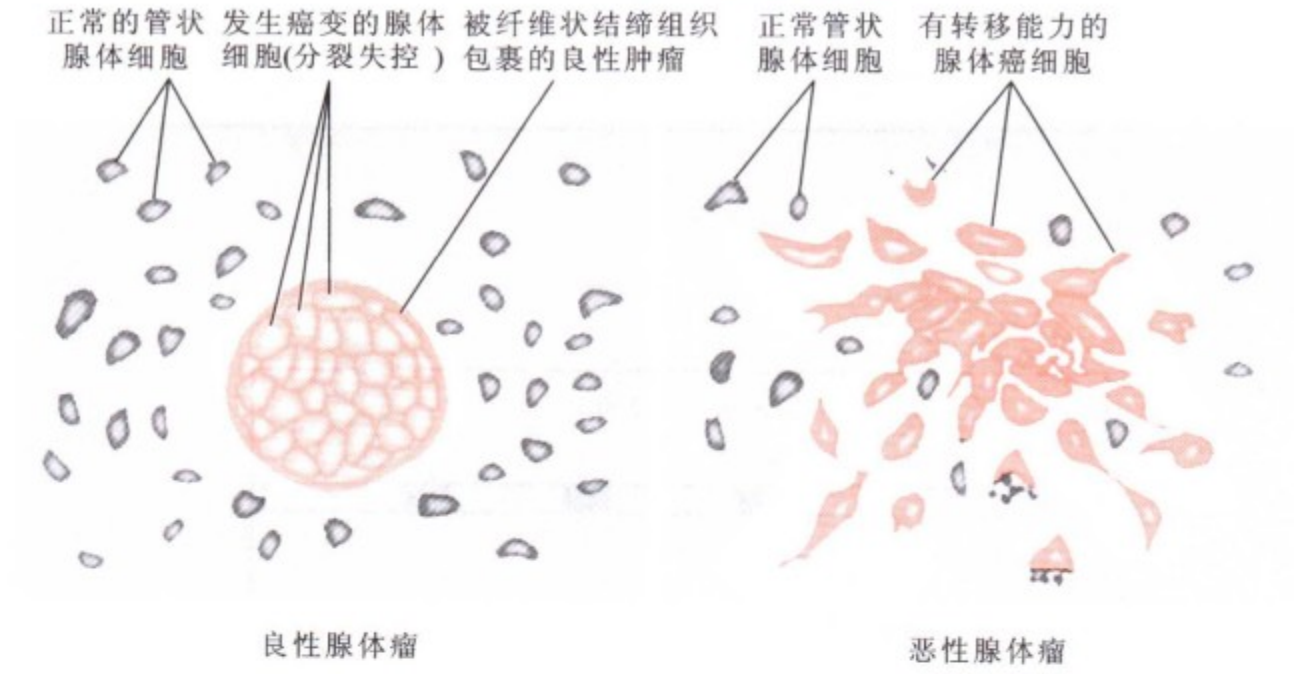



图 9-1 肿瘤组织示意图

界,有相当高的保守性,属于“持家基因”,正常表达时对细胞的生长和分化有调控的作用。

### 9.1.1 反转录病毒致癌基因

早在 1910 年,Rous 就发现带有肉瘤病毒(一种反转录病毒)的鸡肉瘤无细胞滤液能在鸡体内诱发新的肉瘤。这种病毒的基因组是由单链 RNA 组成的,一般只有 6~9 kb。RNA 上的基因数目很少,它们被包裹在由 *gag* 和 *env* 两个基因编码的蛋白质外壳中,其中 *env* 基因指导外壳蛋白的合成,*gag* 基因则指导“鞘”蛋白的合成,这些“鞘”蛋白好像“道钉”一样“箍”在外壳的表面,维持外壳蛋白结构的稳定性(图 9-2)。反转录病毒的复制是由其自身基因组上 *pol* 基因编码的反转录酶指导完成的。当该病毒感染细胞时,*pol* 基因就被宿主的 RNA 聚合酶Ⅱ转录,产生 *pol* mRNA,并在宿主核糖体上合成包括反转录酶、整合酶等多个病毒复制和整合所需要的酶类(图 9-3)。该反转录酶再以病毒 RNA 为模板,转录出与病毒 RNA 序列互补的单链 DNA 分子,并以此为模板,利用宿主 DNA 聚合酶指导合成第二条 DNA 链,以双链 DNA 形式整合到宿主细胞基因组中

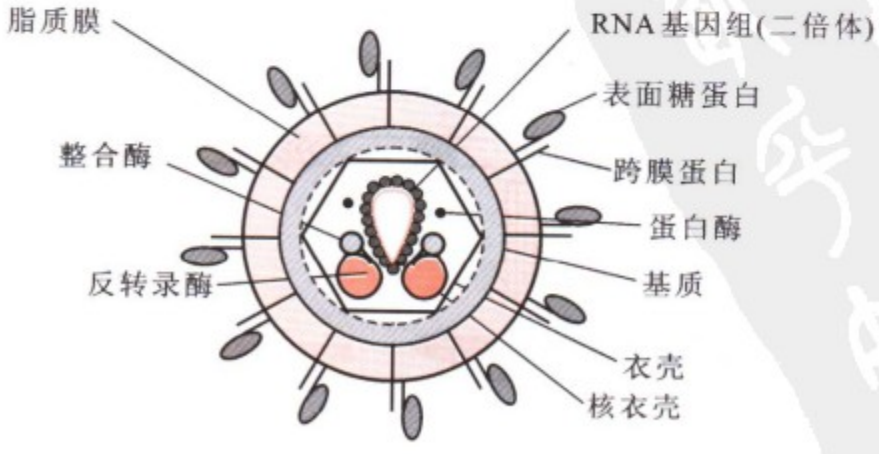



图 9-2 反转录病毒颗粒示意图

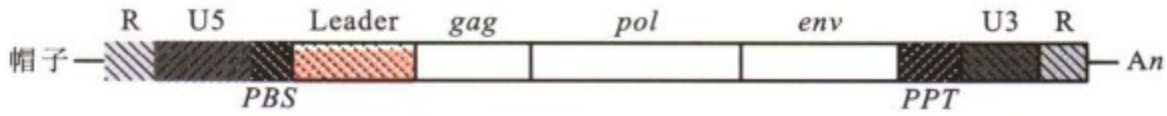



图 9-3 反转录病毒基因结构示意图

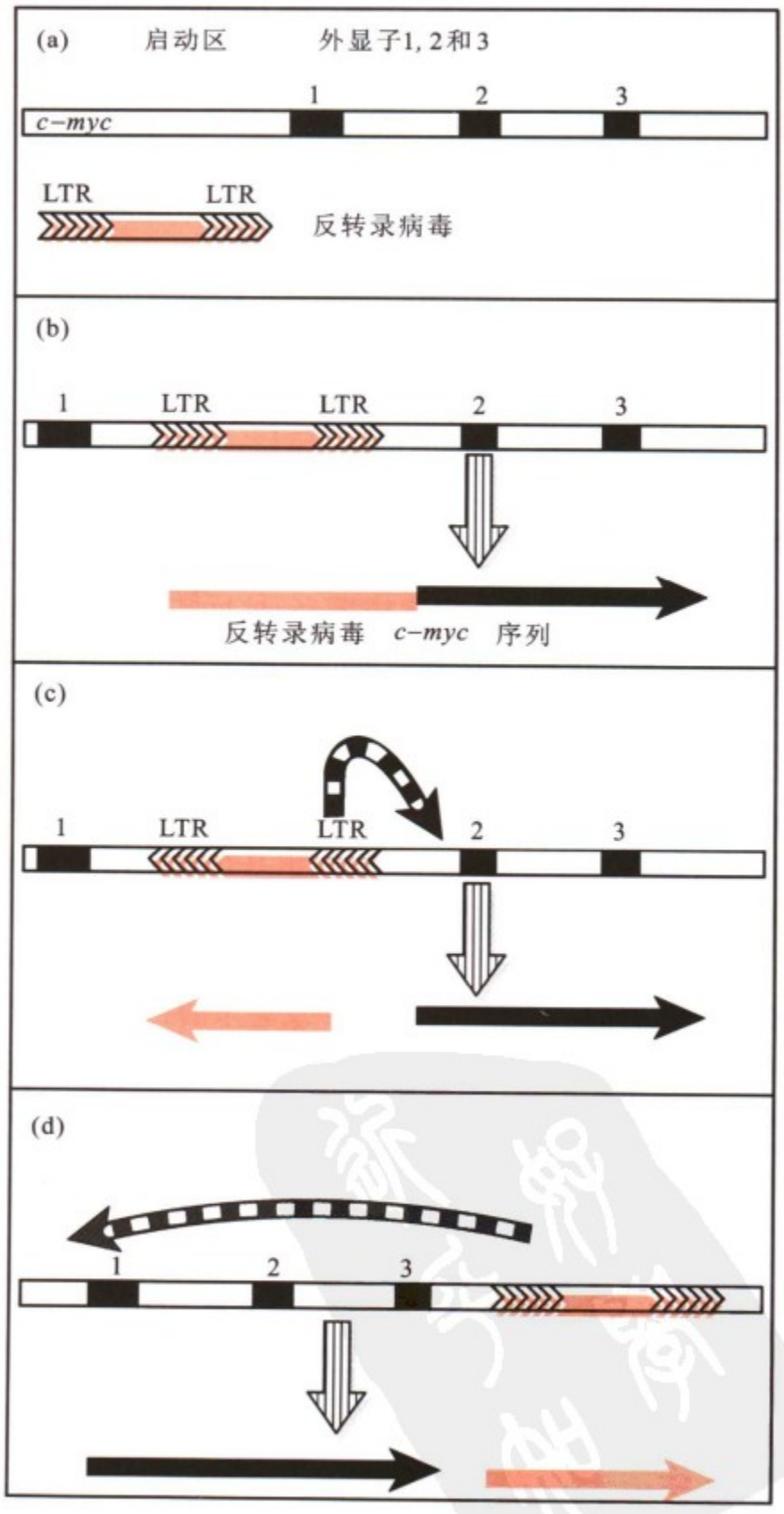



图 9-4 反转录病毒插入引起 *c-myc* 基因活化的几种可能途径

(图 9-4)。被整合的反转录病毒 DNA 分子称为原病毒(provirus),它能指导病毒 mRNA 的合成,并利用宿主细胞中的蛋白质合成器,翻译生成病毒外壳蛋白等,最后组装成病毒颗粒。已经在各种反转录病毒中发现了许多致癌基因,它们感染后诱导宿主产生肿瘤的主要原因是激活特定基因表达,从而破坏宿主细胞本身固有的平衡,导致细胞发生转化。

最早研究的致癌基因是劳斯(Rous)肉瘤病毒中的 *V-Src* 基因,它编码了一个被称为 $p60^{Src}$ 的蛋白质。序列分析表明,该蛋白有 514 个氨基酸,是一个磷酸化蛋白。其 C 端 250 个氨基酸是活性区域,负责将磷酸基团转移到酪氨酸残基上。它的作用主要是使多个靶蛋白发生磷酸化,从而影响它们的功能,加速细胞癌变的过程。

那么,$p60^{Src}$ 使多个蛋白质中的酪氨酸发生磷酸化以后,又是如何促使细胞发生癌变的呢?科学家发现,一种存在于细胞基部质膜附着点内被称为枢纽蛋白(vinculin)的细胞骨架蛋白的磷酸化可能是引起细胞癌变的中心环节。已知质膜附着点的主要作用是通过细胞膜固着蛋白,将细胞固定在所处的表面。它也能使肌动蛋白的纤丝依附在上面。位于附着点内的枢纽蛋白则起着连接肌动蛋白纤丝束和细胞膜固着蛋白的作用。正常情况下,该蛋白上的酪氨酸残基只轻度地磷酸化。而在转化细胞中,由于 $p60^{Src}$ 结合在附着点内靠近或位于枢纽蛋白处,使这一蛋白的磷酸化水平提高了 20 倍以上,明显降低了枢纽蛋白连接肌动蛋白纤丝束和细胞膜固着蛋白的功能,导致肌动蛋白纤丝束松散,细胞黏附能力减弱,容易发生脱落和转移。

1. 急性转化型和非急性转化型

根据反转录病毒转化细胞的能力,可将其分为两大类:一类为急性转化型,另一类则称为非急性转化型。急性转化型反转录病毒具有 3 个主要特征:①这类病毒感染动物后,很短时期内(几天或几周)就出现实体瘤或白血瘤(leukemia);②它们所带的癌基因一般位于病毒基因组内部,也可位于基因组的 3′端,但不会插入结构基因内部;③具有体外转化细胞的能力。非急性转化型则正好相反,它们感染宿主细胞后需要较长时间(几个月,数年甚至数十年)才会致癌。

2. *V-onc* 基因的起源

研究发现,反转录病毒基因组中所带有的 *onc* 基因并非来自病毒本身,而是这些病毒在感染动物或人体之后获得的细胞原癌基因(图 9-5)。这些动物或人原癌基因经病毒修饰和改造后,成为病毒基因组的一部分并具有了致癌性,其作用的靶分子也往往发生了改变。已经证实,各种脊椎动物染色体 DNA 上都有与 *Src* 相类似的 DNA 序列,说明它是正常细胞所固有的。因为正常细胞中这些基因是不表达的,只有在细胞发生癌变时才有活性,所以称为原癌基因。表 9-1 是部分 *c-onc* 与 *v-onc* 序列的比较,从表中不难看出,除了 *erbB* 基因之外,两组基因编码区的相似性非常高。一般说来,细胞原癌基因大都是断裂基因,带有插入序列,而病毒的致癌基因往往是一个完整的没有断裂的读码框。

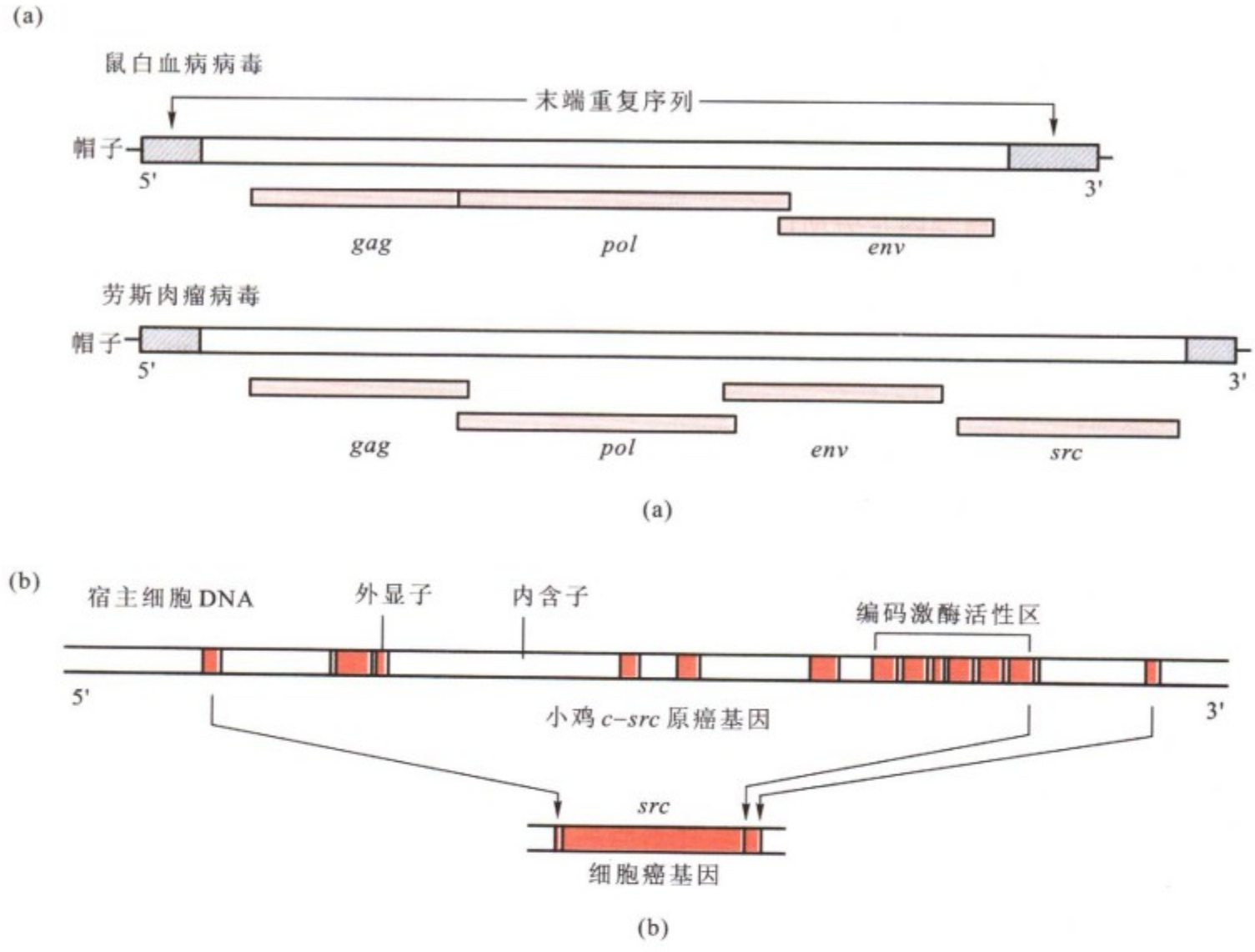



**图 9-5 劳斯肉瘤病毒基因结构及 *c–Src* 原癌基因的转变** (a)鼠白血病病毒(上)和劳斯肉瘤病毒(下)基因组及所编码的主要蛋白质;(b)病毒感染后引起宿主细胞原癌基因转变成癌基因。

**表 9-1 部分 *c-onc* 和 *v-onc* 基因编码区比较**

| 基因 | *c-onc* 中的密码子数 | *v-onc* 中的密码子数 | 不相同氨基酸数 | 相似性/% | *v-onc* 中缺失的区域 |
|---|---|---|---|---|---|
| *Mos* | 369 | 369 | 11 | 97 | 无 |
| *Ha-ras* | 189 | 189 | 3 | 98 | 无 |
| *K-ras* | 189 | 189 | 7 | 96 | 无 |
| *Sis* | 220 | 220 | 18 | 92 | 无 |
| *Myc* | 417 | 417 | 2 | 99 | 无 |
| *Src* | 533 | 514 | 16 | 97 | C 端 |
| *Fms* | 980 | 930 | 20 | 99 | C 端 |
| *ErbB* | 1 210 | 600 | 99 | 83 | N 端、C 端 |
| *ErbA* | 408 | 396 | 22 | 95 | N 端 |
| *Myb* | 640 | 372 | 11 | 97 | N 端、C 端 |

目前已鉴定的致癌基因超过了 100 个，最常见的是 *ras* 与 *myc* 基因家族。突变的 *Ha-ras* 基因产物与结肠癌、肺癌、胰腺癌等相关，*K-ras* 与恶性骨髓瘤、成淋巴细胞瘤等有关，而 *N-ras* 则多见于泌尿系统癌症。研究发现，*c-myc* 表达时相主要在 $G_0/G_1$ 交界或 $G_1$ 早期，而 *C-H-ras* 和 *C-K-ras* 则主要在 $G_1$ 中晚期。有人认为，细胞周期中的 $G_1$ 期对于细胞生长的控制具有关键性意义，因为该时相中存在一个“生长控制点”，在正常情况下，细胞可根据对信号的应答，发挥“生长控制点”的调节作用，使细胞通过该点，进入 S、M 期再回到 $G_1$ 期，或者阻止细胞通过该控制点而使其进入 $G_0$ 期。*c-myc* 的表达可能使细胞获得了通过“生长控制点”的潜能，而 *c-ras* 的表达似乎加强了 *c-myc* 的这种作用潜能，两者协同作用使细胞向 S 期过渡成为可能。也就是说，*c-myc* 和 *c-ras* 表达的协同作用是细胞分裂的必要步骤。实验表明，在肿瘤细胞中，“生长控制点”不起作用，所以瘤细胞一直处于细胞周期循环之中(图 9-6)。

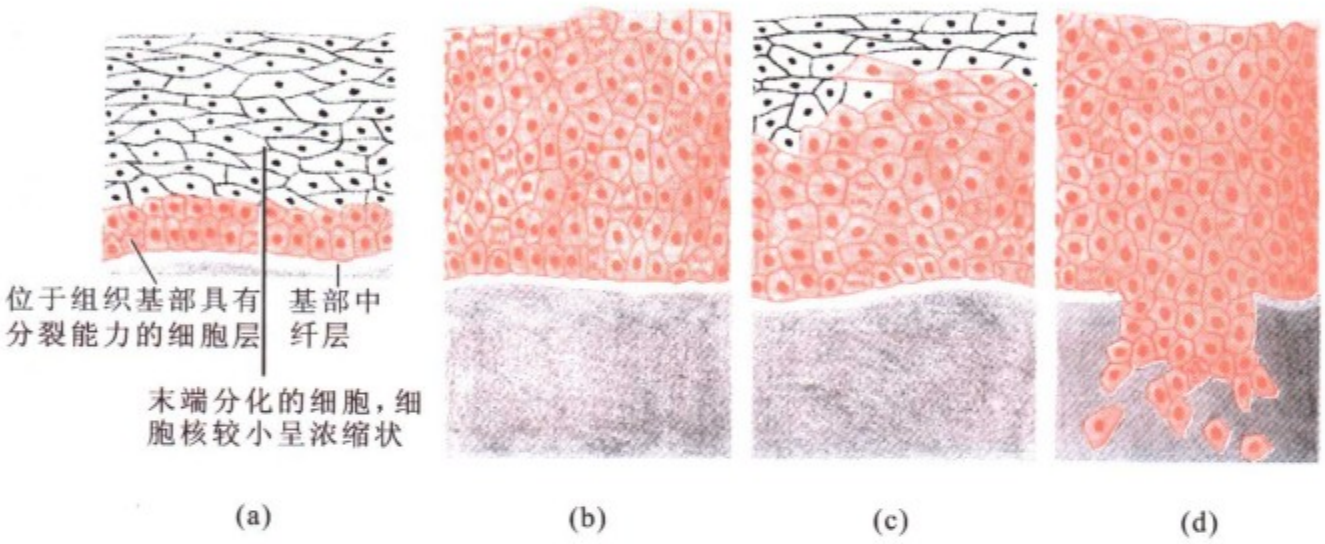
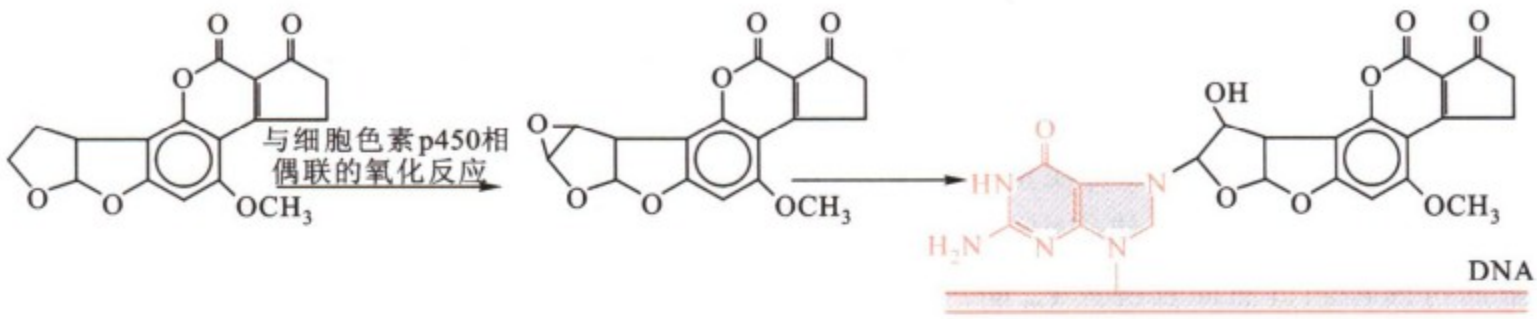



**图 9-6 表皮癌的发生过程示意图** (a)正常表皮组织；(b)固定在原位的表皮组织；(c)发生异常的癌变组织；(d)已经发生转移的恶性肿瘤组织。

## 9.1.2 原癌基因(细胞转化基因)产物及其分类

除了病毒感染会诱发细胞癌变外，尚有许多非病毒因子(如放射性物质、化学试剂亚硝酸、烷化剂等)也能诱导细胞转化。这些因子并没有把致癌基因或其他致癌的遗传信息带入细胞，而仅仅通过某种激活机制改变了细胞内原有的遗传信息，使细胞发生恶性转化。对大量致癌因子的研究证明，它们的作用都是使基因发生突变(图 9-7)，充分说明内源基因在细胞转化中的地位。实验证明，此时，引起细胞恶性转化的因子仍然是 DNA，化学诱变后的细胞中同样存在着决定细胞命运的致癌基因。从人膀胱癌 T24 肿瘤细胞中提取 DNA，转染正常小鼠成纤维细胞使之变为转化细胞。从发生转化的成纤维细胞集落中分离单细胞，培养、扩增后再提取 DNA，第二次转染正常小鼠成纤维细胞，又出现转化细胞集落。如果细胞转化特性是由散布在不同 DNA 片段上的遗传因子决定的，那么，在这个基因重复导入过程中细胞的转化特性就可能消失。事实上，经过连续多

次基因导入后，细胞仍被转化，说明致癌因子不是分布在不同的 DNA 片段上的，而是存在于某个特定的区域。




黄曲霉素　　黄曲霉素-2，3环氧化物　　致癌因子与DNA中的鸟苷酸残基相结合

图 9-7　黄曲霉素(aflatoxin)导致细胞发生癌变的分子机制

根据原癌基因产物在细胞中的位置，可将其分为 3 类：第一类是与膜结合的蛋白，主要有 *erbB*、*neu*、*fms*、*mas* 和 *Src* 基因产物；第二类是可溶性蛋白，包括 *mos*、*sis* 和 *fps* 基因产物；第三类是核蛋白，包括 *myc*、*ets*、*jun* 和 *myb* 等。根据这些蛋白的功能，它们又常被分为 6 类，即蛋白激酶类、生长因子类、生长因子受体类，GTP 结合蛋白类、核蛋白类和功能未知类(表 9-2)。原癌基因产物的一个

表 9-2　原癌基因及其产物的分类

| 类　型 | 相对分子质量 | 产物性质 | 在细胞内定位 |
|---|---|---|---|
| 生长因子类 | | | |
| sis（PDGFβ） | p28 | 与血小板衍生生长因子 β 链有相似性 | 胞质 |
| Fgf－5 | p32 | 与成纤维细胞生长因子相似 | 胞质 |
| 生长因子受体类 | | | |
| Neu(erb B2) | p185 | 与表皮生长因子受体相似 | 质膜 |
| Kit | p145 | 与肥大细胞生长因子受体相似 | 质膜 |
| Mas－1 | | 血管紧张素受体的亚型 | 质膜 |
| Ros－1 | p259 | 具有酪氨酸蛋白激酶活性 | 质膜 |
| Trk | p140 | 原肌球蛋白受体，有激酶活性 | 质膜 |
| Fms | p165 | 与 CSF 受体有相似性 | 胞质/质膜 |
| Erb B | p175 | 与表皮生长因子受体有相似性 | 胞质/质膜 |
| 酪氨酸蛋白激酶(非受体)类 | | | |
| Src 家族 | | | |
| Src | p60 | 具有酪氨酸蛋白激酶活性 | 胞质/质膜 |
| Fyn | p59 | | 胞质/质膜 |
| Syn | | | 胞质/质膜 |
| Lyn | p57 | | 胞质/质膜 |
| Sea | | | 胞质/质膜 |
| Lck | p56 | | 胞质/质膜 |

续表

| 类　型 | 相对分子质量 | 产物性质 | 在细胞内定位 |
|---|---|---|---|
| Hck | p59 | | 胞质/质膜 |
| Ros | p280 | | 胞质/质膜 |
| Abl | p145 | | 胞质/质膜 |
| Fes | p92 | | 胞质/质膜 |
| Yes | p62 | | 胞质/质膜 |
| Fgr | p55 | | 胞质/质膜 |
| Tkl | p35 | | 胞质/质膜 |
| Ret | p135 | | 胞质/质膜 |
| Rel | p68 | | 胞质 |
| 丝/苏氨酸蛋白激酶类 | | | |
| Mos | p43 | 具有丝/苏氨酸蛋白激酶活性 | 胞质 |
| Raf | p73 | 具有丝/苏氨酸蛋白激酶活性 | 胞质 |
| Pim - 1 | p357 | 具有丝/苏氨酸蛋白激酶活性 | 胞质 |
| Raf - 1 | p69 | 具有丝/苏氨酸蛋白激酶活性 | 胞质 |
| Mht | | 与 Raf 有相似性 | 胞质 |
| G 蛋白类 - Ras 家族 | | | |
| H - ras | p21 | 具有 GTP 结合作用和 GTP 酶活性 | 质膜内面 |
| K - ras | p21 | 具有 GTP 结合作用和 GTP 酶活性 | |
| N - ras | p21 | 具有 GTP 结合作用和 GTP 酶活性 | |
| 核蛋白类 | | | |
| Ets 家族 | | | |
| ets - 1 | p51 | DNA 结合蛋白 | 细胞核 |
| ets - 2 | p56 | DNA 结合蛋白 | 细胞核 |
| elk - 1 | | DNA 结合蛋白 | 细胞核 |
| Jun 家族 | | | |
| Jun | p47 | DNA 结合蛋白 | 细胞核 |
| jun B | | DNA 结合蛋白 | 细胞核 |
| jun D | | DNA 结合蛋白 | 细胞核 |
| Fos 家族 | | | |
| Fos | p62 | DNA 结合蛋白 | 细胞核 |
| fos b | p45 | DNA 结合蛋白 | 细胞核 |
| fra - 1 | p38 | DNA 结合蛋白 | 细胞核 |
| Myc 家族 | | | |
| Myc | p62 | DNA 结合蛋白 | 细胞核 |
| N - myc | p62 | DNA 结合蛋白 | 细胞核 |
| L - myc | p60 | DNA 结合蛋白 | 细胞核 |
| Myb | P75 | DNA 结合蛋白 | 细胞核 |
| Ski | p60 | DNA 结合蛋白 | 细胞核 |

注：功能未知的：bcl - 1，bcl - 2，vav，evi - 1，spi - 1，gsp，ove，crk，mpl，rap 1A，rab 4，arh - 6，rabl，rab2

共同特征是它们都能诱发一系列与细胞生长分化有关的基因表达,从而改变细胞的表型。

### 9.1.3 原癌基因的表达调控

原癌基因在正常细胞中通常以单拷贝形式存在,只有低水平的表达或根本不表达。在很多情况下,原癌基因的结构发生了点突变或插入、重排、缺失及扩增等,改变其转录活性(图 9-8)。

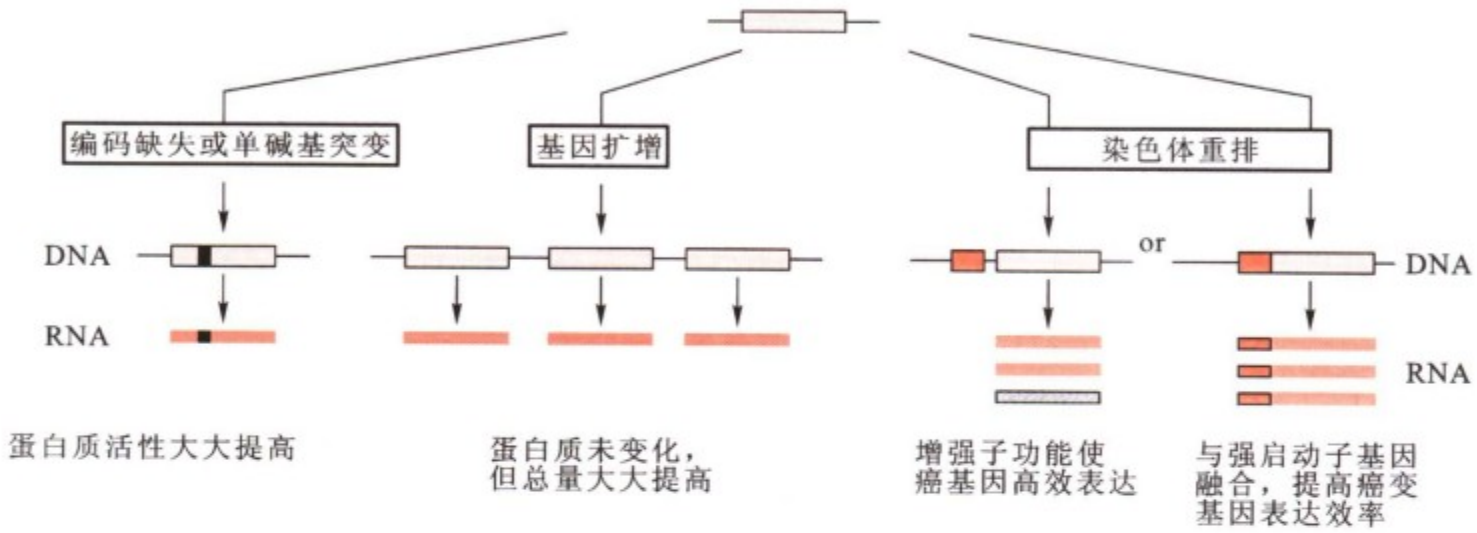



图 9-8 细胞中原癌基因转变为癌基因的主要途径

1. 点突变

虽然原癌基因活化导致癌变的原因多种多样,但点突变可能是最常见的机制。在 *ras* 基因家族中,存在着许多原癌基因发生点突变的实例。*Ha-ras* 和 *Ki-ras* 基因最初是从小鼠肉瘤病毒中分离出来的,*N-ras* 是从人神经母细胞瘤细胞系中分离得到的,*C-ras* 基因存在于非肿瘤细胞生物机体中,可能是细胞发挥正常的生理功能所必需的。研究发现,*ras* 基因编码了一个相对分子质量为 $2.1\times10^4$ 癌蛋白 (p21),从人类膀胱癌细胞系 T24 DNA 中克隆的 *Ha-ras* 基因能够诱发 NIH/3T3 细胞转化,而从正常细胞 DNA 中克隆的该原癌基因没有这种功能。因为 T24 膀胱癌细胞中 *Ha-ras* 基因的表达水平较正常细胞并无明显提高,科学家推测点突变使癌基因产物 p21 蛋白的结构和功能发生了某些变化(图 9-9)。如人类肺癌细胞系 Hs242 的转化基因与 *Ha-ras* 高度相似,在这个基因中导致转化活性的遗传损伤是第二个外显子中引起 p21 蛋白第 61 位谷氨酰胺被亮氨酸所替代的一个点突变。所以,p21 分子某些部位发生单个氨基酸替代足以引起蛋白质构象的改变,并使细胞获得转化活性。

在肺癌和结肠癌细胞系中也发现了 *Ki-ras* 型转化基因。肺癌细胞系含有第 12 位密码子由 TGT 取代 GGC,从而以半胱氨酸取代甘氨酸的 *Ki-ras*2 转化基因;而 SW480 结肠癌细胞系也含有 *Ki-ras* 转化基因,其第 12 位密码子为 GTT,编码了缬氨酸。这两个细胞系都经历了点突变激活过程,突变部位与人类膀胱癌细胞系 T24 *Ha-ras* 的对应部位相同。现在已经证明由 NIH/3T3 细胞转染检出的 *ras* 原癌基因中,第 12 位密码子是个突变“热点”。

| | 第一外显子 | | | 第二外显子 | | | 转化活性 |
|---|---|---|---|---|---|---|---|
| | 11 | 12 | 13 | 60 | 61 | 62 | |
| 原C-H-*ras* | …GCC<br>丙氨酸 | …GCC<br>丙氨酸 | GGT…<br>甘氨酸 | …GGC<br>甘氨酸 | CAG<br>谷氨酰胺 | GAG…<br>谷氨酸 | − |
| 膀胱癌<br>C-H-*ras* | …GCC<br>丙氨酸 | GTC<br>缬氨酸 | GGT…<br>甘氨酸 | …GGC<br>甘氨酸 | CAG<br>谷氨酰胺 | GAG…<br>谷氨酸 | + |
| 肺癌C-H-*ras* | …GCC<br>丙氨酸 | GTC<br>缬氨酸 | GGT…<br>甘氨酸 | …GGC<br>甘氨酸 | CTG<br>亮氨酸 | GAG…<br>谷氨酸 | + |
| MSV-H-*ras* | …GCC<br>丙氨酸 | AGA<br>精氨酸 | GGT…<br>甘氨酸 | …GGC<br>甘氨酸 | CTG<br>亮氨酸 | GAG…<br>谷氨酸 | + |
| 乳腺癌<br>NMU-H-*ras* | …GCC<br>丙氨酸 | GAA<br>谷氨酸 | GGT…<br>甘氨酸 | …GGC<br>甘氨酸 | ?<br>? | GAG…<br>谷氨酸 | + |
| 原C-K-*ras* | …GCT<br>丙氨酸 | GGT<br>甘氨酸 | GGC…<br>甘氨酸 | …GGT<br>甘氨酸 | CAA<br>谷氨酰胺 | GAA…<br>谷氨酸 | − |
| 肺癌C-K-*ras* | …GCT<br>丙氨酸 | TGT<br>半胱氨酸 | GGC…<br>甘氨酸 | …GGT<br>甘氨酸 | CAA<br>谷氨酰胺 | GAA…<br>谷氨酸 | + |
| 大肠癌<br>C-K-*ras* | …GCT<br>丙氨酸 | GTT<br>缬氨酸 | GGC…<br>甘氨酸 | …GGT<br>甘氨酸 | CAA<br>谷氨酰胺 | GAA…<br>谷氨酸 | + |
| MSV-K-*ras* | …GCT<br>丙氨酸 | AGT<br>丝氨酸 | GGC…<br>甘氨酸 | …GGT<br>甘氨酸 | CAA<br>谷氨酰胺 | GAA…<br>谷氨酸 | + |

▌图 9-9 *ras* 基因的点突变及转化活性分析

2. LTR 插入

LTR 是反转录病毒基因组两端的长末端重复序列(long terminal repeat),其结构中含有强的启动子序列，当 LTR 插入原癌基因启动子区域或邻近部位后，可从根本上改变基因的正常调控规律。最早发现 LTR 插入激活现象是在慢性反转录病毒感染的细胞中。LTR 插入到 *c-myc* 5′上游启动子附近,使 *c-myc* 的转录水平大大增加。其他原癌基因中也发现了类似现象。LTR 插入后,与血小板衍生生长因子 β 链(PDGF-β)相似的 *c-sis* 基因的转录和翻译水平均有明显增加。表 9-3 是受 LTR 插入激活的部分原癌基因。

3. 基因重排

在大多数类型的人肿瘤中存在染色体数目和结构异常现象,说明存在原癌基因重排的可能。现已查明基因重排包括原癌基因之间,及原癌基因与非原癌基因之间的重排。重排后基因内部结构可不受影响,也可能发生改变。在 Burkitt 淋巴瘤中,*Ig* 基因与 *c-myc* 之间发生重排。正常情况下,*c-myc* 定位于 $8q^{24}$,免疫球蛋白重链基因(IgH)定位于 $14q^{32}$,轻链 λ 基因(Igλ)定位于 $22q^{12}$,轻链 κ 基因(Igκ)定位于 $2p^{11}$。重排时,*c-myc* 从其上游区到第 2 外显子的区域内断裂,易位至 IgH、Igκ 或 Igλ 的位点，使 *Ig* 基因与 *c-myc* 相连在一起,*Ig* 基因 5′端的启动子发挥作用,使原来不表达的 *c-myc* 大量表达。由于人体的特殊需求,通常 *Ig* 基因的表达能力非常强。*c-myc* 一旦与 *Ig* 基因启动子和增强子重组后,可使原

表 9-3 LTR 插入激活的原癌基因

| 原癌基因 | LTR 来源 | 诱发的肿瘤 |
|---|---|---|
| | 禽类白血病病毒 | 禽类黏液性囊性淋巴瘤 |
| *Myc* | 鼠白血病病毒 | 鼠 T－细胞淋巴瘤 |
| | 猫白血病病毒 | 猫 T－细胞淋巴瘤 |
| *Erb B* | 禽类白血病病毒 | 禽类红白血病 |
| *H-ras* | 禽类白血病病毒 | 禽肾细胞瘤 |
| *K-ras* | 小鼠白血病病毒 | 小鼠髓性细胞白血病 |
| *Myb* | 禽类白血病病毒 | 禽类黏液性囊性淋巴瘤 |
| *Fms* | 小鼠白血病病毒 | 小鼠髓性细胞白血病 |
| *Int* －1 | 小鼠乳头瘤病毒 | 小鼠乳腺癌 |
| *Pim* －1 | 小鼠白血病病毒 | 小鼠 T－细胞淋巴瘤 |
| *Lck* | 小鼠白血病病毒 | 小鼠 T－细胞淋巴瘤 |
| *evi* －1 | 小鼠白血病病毒 | 小鼠髓性细胞白血病 |

来的调控机制失去作用，而导致极高水平的表达。在正常人体细胞中，非受体型酪氨酸蛋白激酶基因 *abl* 位于第 9 号染色体上，表达量极低，不会诱发癌变。在慢性骨髓瘤患者细胞中，该基因却被转移到第 22 号染色体上，与 *bcr* 基因融合，表达量大为提高，导致细胞分裂失控，发生癌变(图 9-10)。

4. 缺失

很多原癌基因 5′上游区存在负调控序列，一旦该序列发生缺失或突变，就

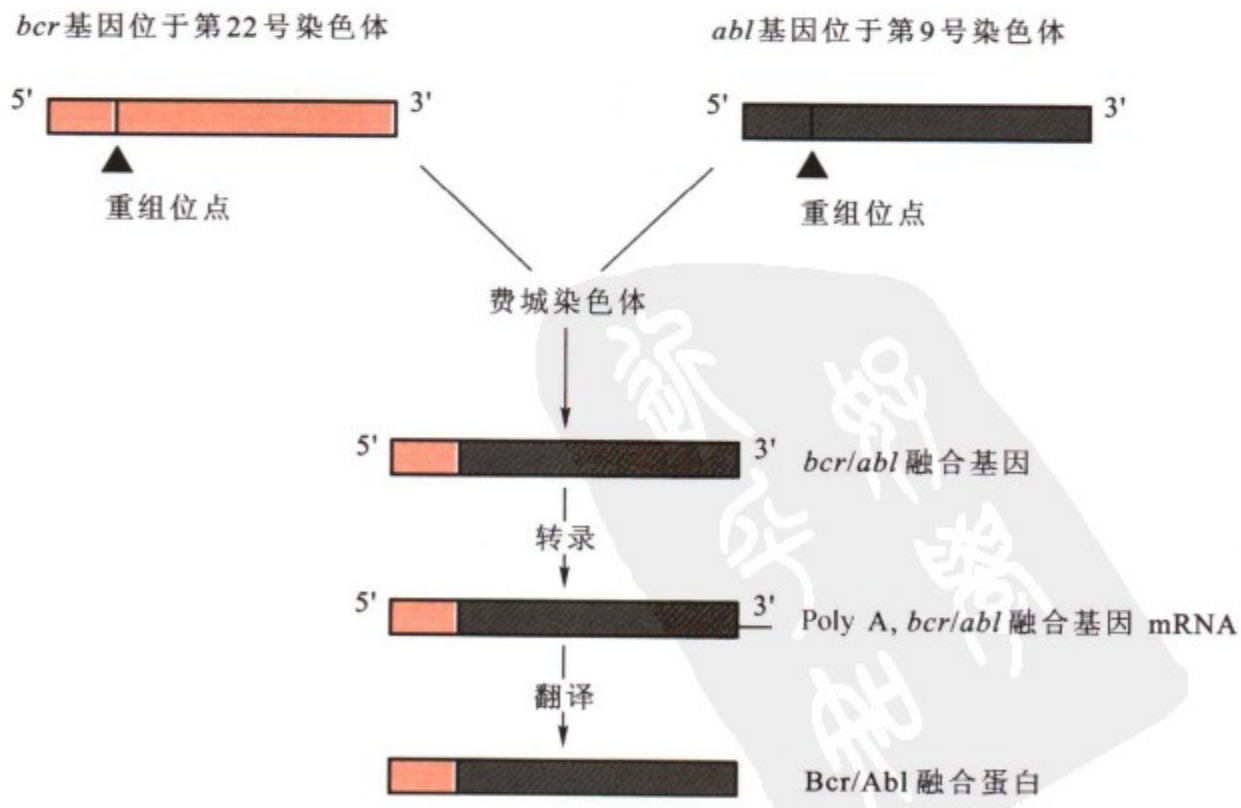



图 9-10 *abl* 原癌基因通过选择性染色体重排转变成细胞癌基因

| | 第一外显子 | | | 第二外显子 | | | 转化活性 |
|---|---|---|---|---|---|---|---|
| | 11 | 12 | 13 | 60 | 61 | 62 | |
| 原C-H-*ras* | …GCC | …GCC | GGT… | …GGC | CAG | GAG… | |
| | 丙氨酸 | 丙氨酸 | 甘氨酸 | 甘氨酸 | 谷氨酰胺 | 谷氨酸 | − |
| 膀胱癌 | …GCC | GTC | GGT… | …GGC | CAG | GAG… | |
| C-H-*ras* | 丙氨酸 | 缬氨酸 | 甘氨酸 | 甘氨酸 | 谷氨酰胺 | 谷氨酸 | + |
| 肺癌C-H-*ras* | …GCC | GTC | GGT… | …GGC | CTG | GAG… | |
| | 丙氨酸 | 缬氨酸 | 甘氨酸 | 甘氨酸 | 亮氨酸 | 谷氨酸 | + |
| MSV-H-*ras* | …GCC | AGA | GGT… | …GGC | CTG | GAG… | |
| | 丙氨酸 | 精氨酸 | 甘氨酸 | 甘氨酸 | 亮氨酸 | 谷氨酸 | + |
| 乳腺癌 | …GCC | GAA | GGT… | …GGC | ? | GAG… | |
| NMU-H-*ras* | 丙氨酸 | 谷氨酸 | 甘氨酸 | 甘氨酸 | ? | 谷氨酸 | + |
| 原C-K-*ras* | …GCT | GGT | GGC… | …GGT | CAA | GAA… | |
| | 丙氨酸 | 甘氨酸 | 甘氨酸 | 甘氨酸 | 谷氨酰胺 | 谷氨酸 | − |
| 肺癌C-K-*ras* | …GCT | TGT | GGC… | …GGT | CAA | GAA… | |
| | 丙氨酸 | 半胱氨酸 | 甘氨酸 | 甘氨酸 | 谷氨酰胺 | 谷氨酸 | + |
| 大肠癌 | …GCT | GTT | GGC… | …GGT | CAA | GAA… | |
| C-K-*ras* | 丙氨酸 | 缬氨酸 | 甘氨酸 | 甘氨酸 | 谷氨酰胺 | 谷氨酸 | + |
| MSV-K-*ras* | …GCT | AGT | GGC… | …GGT | CAA | GAA… | |
| | 丙氨酸 | 丝氨酸 | 甘氨酸 | 甘氨酸 | 谷氨酰胺 | 谷氨酸 | + |

▌图 9-9 *ras* 基因的点突变及转化活性分析

2. LTR 插入

LTR 是反转录病毒基因组两端的长末端重复序列(long terminal repeat),其结构中含有强的启动子序列，当 LTR 插入原癌基因启动子区域或邻近部位后，可从根本上改变基因的正常调控规律。最早发现 LTR 插入激活现象是在慢性反转录病毒感染的细胞中。LTR 插入到 *c-myc* 5′上游启动子附近,使 *c-myc* 的转录水平大大增加。其他原癌基因中也发现了类似现象。LTR 插入后,与血小板衍生生长因子 β 链(PDGF-β)相似的 *c-sis* 基因的转录和翻译水平均有明显增加。表 9-3 是受 LTR 插入激活的部分原癌基因。

3. 基因重排

在大多数类型的人肿瘤中存在染色体数目和结构异常现象,说明存在原癌基因重排的可能。现已查明基因重排包括原癌基因之间,及原癌基因与非原癌基因之间的重排。重排后基因内部结构可不受影响,也可能发生改变。在 Burkitt 淋巴瘤中,*Ig* 基因与 *c-myc* 之间发生重排。正常情况下,*c-myc* 定位于 $8q^{24}$,免疫球蛋白重链基因(IgH)定位于 $14q^{32}$,轻链 λ 基因(Igλ)定位于 $22q^{12}$,轻链 κ 基因(Igκ)定位于 $2p^{11}$。重排时,*c-myc* 从其上游区到第 2 外显子的区域内断裂,易位至 IgH、Igκ 或 Igλ 的位点，使 *Ig* 基因与 *c-myc* 相连在一起,*Ig* 基因 5′端的启动子发挥作用,使原来不表达的 *c-myc* 大量表达。由于人体的特殊需求,通常 *Ig* 基因的表达能力非常强。*c-myc* 一旦与 *Ig* 基因启动子和增强子重组后,可使原

递生长刺激信号的部位有 3 处:①癌基因产物本身模拟了生长因子,因而与相应的受体作用,以自分泌的方式刺激细胞生长;②癌基因产物模拟了已结合配体的生长因子受体,从而在无外源生长因子时提供促进细胞分裂的信号;③癌基因产物作用于细胞内生长控制途径,解除此途径对外源刺激信号的需求。人血小板衍生生长因子(PDGF)与猴肉瘤病毒(SSV)的 *V-ras* 癌基因产物,上皮生长因子(EGF)与 *Src* 及 *V-erb* 癌基因产物之间都存在着极高的相似性,表明生长因子与癌基因转化有关。

3. 抑癌基因产物对原癌基因的调控

因为抑癌基因产物能够抑制细胞的恶性增殖,所以它也被认为是一种隐性癌基因。当细胞内由于某种原因造成这些基因的表达受抑制时,原癌基因就活跃表达,引起细胞癌变。抑癌基因与原癌基因在生物学上有以下几点差异:①在功能上,抑癌基因在细胞生长中起负调控作用,抑制增殖,促进分化、成熟和衰老,引导多余的细胞进入程序化死亡途径,而原癌基因的作用则正好相反;②在遗传方式上,原癌基因是显性的,激活后即参与促进细胞增殖和癌变过程,而抑癌基因在细胞水平上是隐性的,因为抑癌基因的两个等位基因中失去任何一个都不影响其抑癌功能,只有两个等位基因全部失活才能失去抑癌功能;③在突变发生的细胞类型上,不仅体细胞中可能发生抑癌基因突变,生殖细胞中也可能发生抑癌基因突变并通过生殖细胞得到遗传。原癌基因突变只发生在体细胞中。

*p53* 是通过杂合缺失鉴定的一个抑癌基因,在星形细胞癌、乳腺癌、肺癌、肠癌及骨肉瘤中都有高频率缺失现象。从癌细胞中得到的 *p53* 基因,其保守序列区有单一位点的突变,推测可能由于这一突变导致 *p53* 基因产物结构与功能的改变,失去抑癌活性。另外,在人类某些肿瘤细胞中也可观察到 *p53* 等位基因的缺失,这种缺失同样与肿瘤的发生相关。*p53* 这种抑癌作用可能是通过其蛋白与 DNA 的结合而实现的。这种结合既影响 DNA 的复制,又影响基因的转录。*Rb* 基因是从视网膜纤维瘤中克隆到的另一个抑癌基因,其功能是阻止处于 $G_0/G_1$ 期的细胞进入 S 期,从而控制细胞增殖。正常 *Rb* 基因的表达几乎可抑制所有培养细胞的分裂。

总之,细胞的癌变绝非是单一因素引起的,它可能是显性的癌基因与隐性的抑癌基因经过多阶段的协同作用形成的。一般说来,在一种肿瘤发生过程中起主要作用的几乎不可能是两种以上的癌基因,但很可能涉及两个以上抑癌基因的突变。

4. 外源信号对原癌基因表达的影响

细胞外信号(包括生长因子、激素、神经递质、药物等)作用于靶细胞后,通过细胞膜受体系统或其他直接途径被传递至细胞内,再通过多种蛋白激酶的活化,对转录因子进行修饰,然后引发一系列基因的转录激活(图 9-11)。这一过程进行得很快,通常在几分钟至几十分钟内即可完成。据测定,在这么短时间内,大约有 70~80 种基因的转录被激活,其中包括众多的原癌基因。研究表明,*c-fos* 基因的转录激活可能是最快、最显著的,信号刺激后几分钟内即可检测

到 *c-fos* 基因产物。随后，另一些编码核蛋白的原癌基因如 *c-myc*、*c-myb*、*c-erbB*、*c-ets*等的转录也相继被激活。这些对细胞外信号反应迅速的基因被称为快速反应基因。快速反应基因的转录有两个特点：①不受蛋白质合成抑制剂的阻断。在许多情况下，还会因为蛋白质合成抑制剂而抑制不稳定蛋白质的合成并使转录产物的半衰期显著延长。②快速反应基因的转录激活维持时间很短，通常在半小时内完成，然后逐渐恢复原状。从这一点看，这类基因产物的主要功能是作为转录因子及调节因子去调节别的基因表达，以介导细胞外信号在细胞水平上的应答。

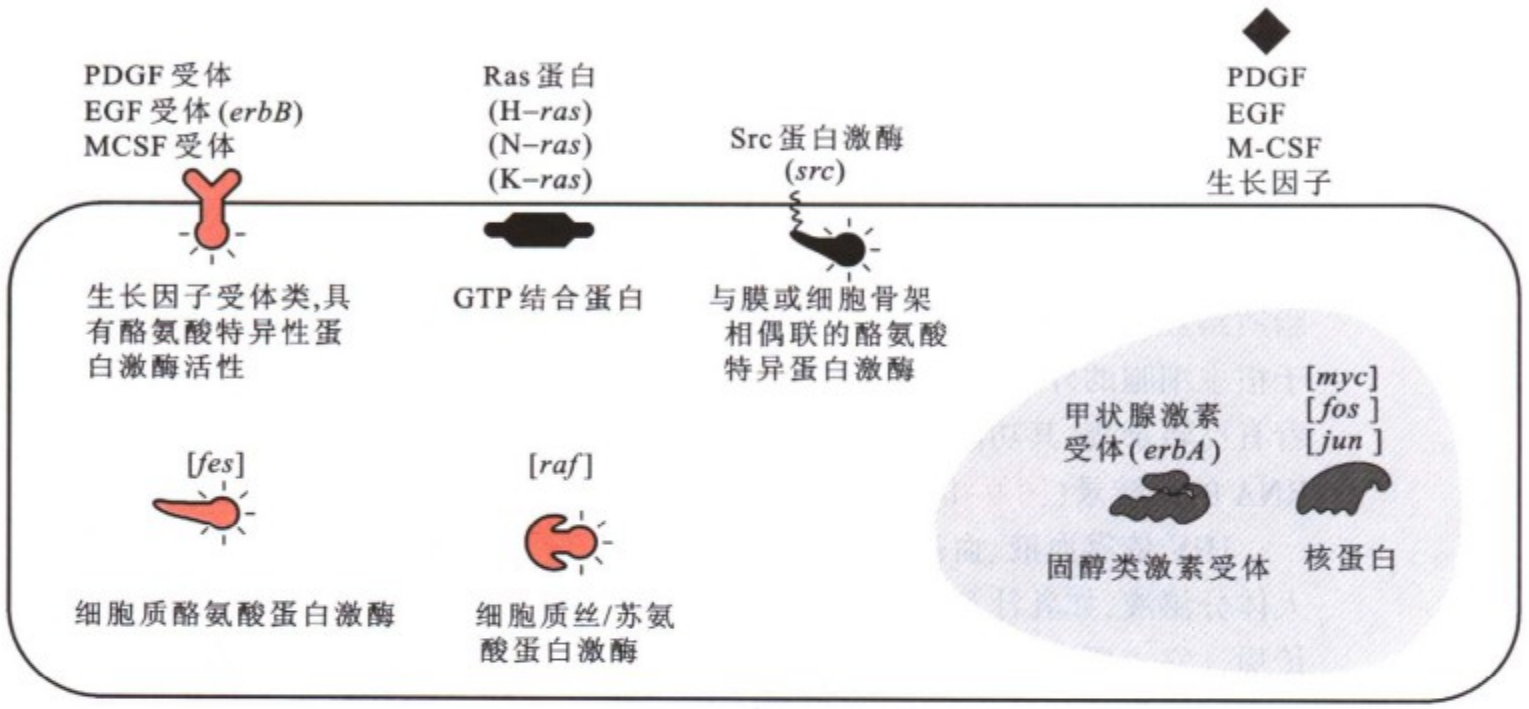



图 9-11 许多原癌基因参与细胞信号转导过程

## 9.2 人类免疫缺陷病毒——HIV

人类免疫缺陷病毒(HIV)，俗称艾滋病(AIDS)病毒，诱发人类获得性免疫缺陷综合征。该病毒在分类上属反转录病毒科(*Retroviridae*)慢病毒属中的灵长类免疫缺陷病毒亚属，过去还有人将其命名为 LAV、ARV、IDAV 和 HTLV3，现统一命名为 HIV。1983 年，法国巴斯德研究所的 Montaginer 和美国国家卫生研究院癌症研究所 Gallo 等人首次证实 HIV 是艾滋病的病因。该病已在二十多年间席卷全球，几乎没有一个国家或地区可以幸免。已发现人类免疫缺陷病毒主要有两种，即 HIV-Ⅰ和 HIV-Ⅱ。HIV-Ⅰ是从欧洲和美洲分离的毒株，与猴艾滋病病毒只有约 45%的相似性，它的致病力很强，是引起全球艾滋病流行的主要病原。HIV-Ⅱ与猴艾滋病病毒的相似性高达 75%，其毒力较弱，引起艾滋病的病程较长，症状较轻，且主要局限于西部非洲(表 9-4)。有关 HIV 的研究主要是针对 HIV-Ⅰ进行的。

表 9-4 反转录病毒科主要成员

| 属 名 | 典型成员 |
|---|---|
| 哺乳动物 C 型反转录病毒属 | 小鼠乳腺瘤病毒 |
| D 型反转录病毒属 | 鼠白血病病毒 |
| 禽 C 型反转录病毒属 | Mason-Pfizer 猴病毒 |
| 泡沫病毒属 | 禽白血病病毒 |
| 人嗜 T－淋巴细胞病毒 | 人泡沫病毒 |
| 白血病病毒属 | 人嗜 T－淋巴细胞病毒 I 型 |
| 慢病毒属 | 人类免疫缺陷病毒 |

## 9.2.1 HIV 病毒颗粒的形态结构及传播

艾滋病病毒粒子是一种直径约为 100 nm 的球状病毒，粒子外包被着由两层脂质组成的脂膜，这种结合有许多糖蛋白分子(主要是 gp41 和 gp120)的脂质源于宿主细胞的外膜。蛋白质 p24 和 p18 组成其核心，内有基因组 RNA 链，链上附着有反转录酶，其功能是催化病毒 RNA 的反转录(图 9-12)。

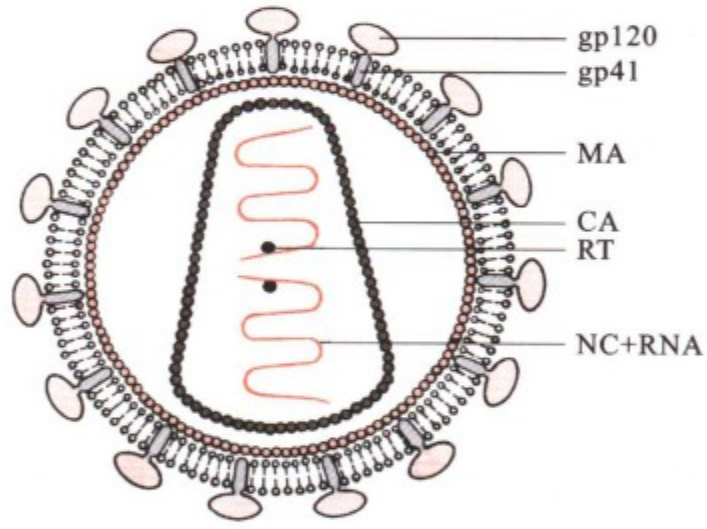



图 9-12 HIV-Ⅰ病毒粒子结构模型图 gp120 和 gp41 是外膜蛋白，MA 是内膜蛋白，CA 是外壳蛋白，NC 是核衣壳蛋白，RT 是转录酶。

HIV 依靠血液、血液制品以及人体分泌液，如乳汁和精液等进行传播，它主要感染 T4-淋巴细胞，也可以感染其他类型如 B-淋巴细胞和单形核细胞等。HIV 感染后可引起明显病变，形成多核巨细胞，并导致细胞死亡。HIV 病毒可以通过所感染细胞扩散到全身，已在淋巴细胞、脑、胸腺、脾等组织发现了该病毒。不同毒株在试管内感染细胞的能力差异很大，说明自然界广泛存在着突变株。迄今为止，尚缺乏感染艾滋病的动物模型。猩猩感染 HIV 后，会产生抗体和暂时性 T 细胞比例失调，但并不发展成艾滋病的临床病症。

## 9.2.2 HIV 基因组及其编码的蛋白

1. *HIV* 基因组结构

*HIV* 基因组由两条单链正链 RNA 组成，每个 RNA 基因组约为 9.7 kb。在 RNA5′端有一帽子结构($m^7G5'GmpNp$)，3′端有 poly A 尾巴。HIV 的主要基因结构和组织形式与其他反转录病毒相同，均由 5′端 LTR、结构蛋白编码区(*gag*)、

蛋白酶编码区(*pro*)、具有多种酶活性的蛋白编码区(*pol*)、外膜蛋白(*env*)和 3′端 LTR 组成。HIV 较其他反转录病毒复杂,因为它有较多的调控基因。为了最大限度地利用有限的基因,*HIV-Ⅰ*基因编码区有很多重叠,尤其在基因组的 3′端。*HIV* 基因组中的部分基因如 *tat* 和 *rev* 是不连续的,被插入的内含子分隔成两个外显子(图 9-13)。*HIV-Ⅰ*至少有 4 个功能性的剪接供体位点和 6 个受体位点。

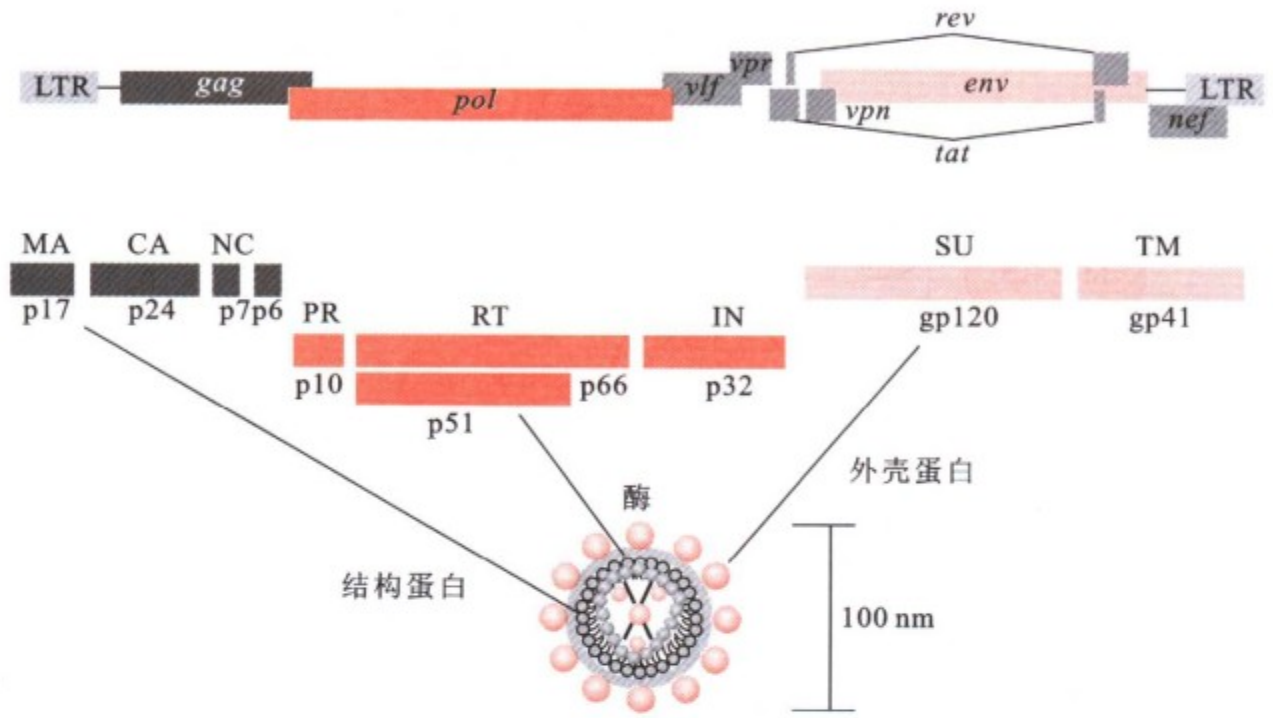



**图 9-13 HIV-Ⅰ基因组结构及所编码的主要蛋白质** *gag* 基因编码了由 p55 前体蛋白切割形成的 MA(p17)、CA(p24)和 NC(p15)等四个结构蛋白(因为 p15 进一步被切割为 p7 和 p6);PR(p10)是蛋白酶,IN(p32)是整合酶,p66 和 p51 是反转录酶;gp120 和 gp41 是主要外壳蛋白。

2. *HIV* 编码的蛋白质及其主要功能

HIV 具有独特的粒子结构,其结构蛋白主要包括 4 个基因。*gag* 基因编码病毒的核心蛋白,翻译时先形成一个 $5.5\times10^4$ 的前体蛋白(p55),然后在 HIV 蛋白酶的作用下被切割成 p17、p24、p15 三个蛋白。p24 和 p17 分别组成 HIV 颗粒的外壳(CA)和内膜蛋白(MA),p15 进一步被切割成与病毒 RNA 结合的核衣壳蛋白(NC)p7 和 p6。*pol* 基因编码病毒复制所需的酶类包括反转录酶 p66、整合酶 p32。从 *pol* 和 *gag* 基因重叠区内起始的一段序列为 *PR* 基因,它编码蛋白酶 p10,在切割上述 HIV 蛋白前体产生成熟蛋白的过程中起主要作用。*env* 基因编码了 $8.8\times10^4$ 的蛋白质,经糖基化后其相对分子质量增至 $1.6\times10^5$,是 HIV 包膜糖蛋白 gp160 的前体。该前体蛋白在蛋白酶作用下被切割成 gp120 和 gp41,gp120 暴露于病毒包膜外,称为外膜蛋白,感染细胞时可与细胞的 CD4 受体蛋白相结合;外膜糖蛋白 gp120 与 CD4 结合后,与 G 蛋白偶联受体(GPCR)家族中的一个跨膜蛋白发生相互作用。gp41 被称为跨膜蛋白(TM),嵌入病毒包膜脂质中。当 gp120 与 CD4 受体结合后,其构象改变导致与 gp41 分离,独立的 gp41 可插入细胞膜,造成膜融合使病毒核心进入细胞内。最新的研究表明,这些辅助受体对 HIV 的侵染和致病是非常重要的,图 9-14 表示 HIV-Ⅰ与 CD4 及辅助

受体 CCR4、CCR5 在感染小神经胶质细胞和巨噬细胞时发生相互作用的情况。表 9-5 是已经发现的辅助受体，表 9-6 则系统介绍了 HIV 编码的主要蛋白及其功能。

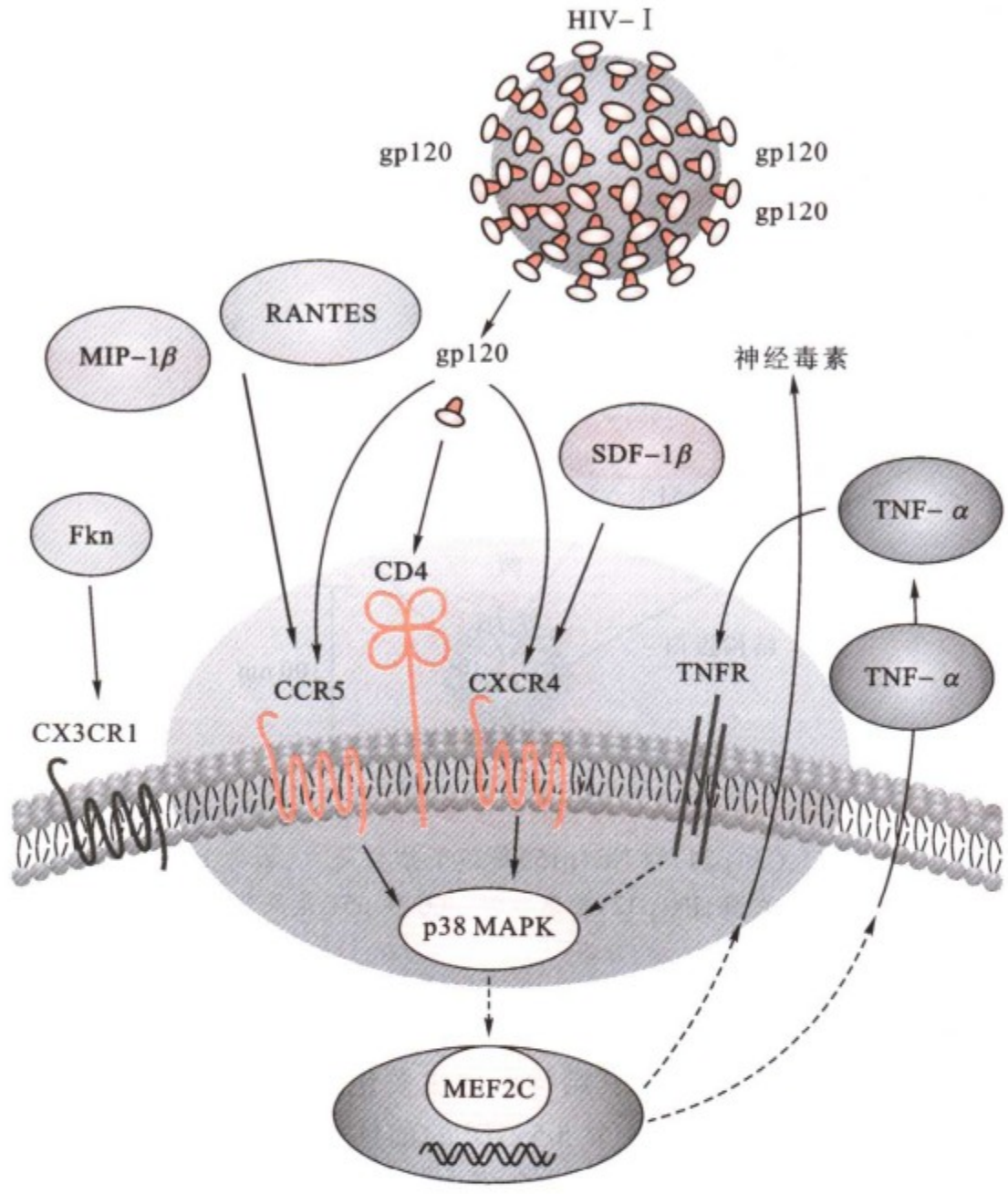



图 9-14 HIV-Ⅰ感染小神经胶质细胞和巨噬细胞过程中信号转导及与其受体 CD4 和辅助受体相互作用示意图

3. HIV 膜蛋白主要功能区

HIV-Ⅰ中的 gp160 由 845~870 个氨基酸残基组成，在第 511 位被蛋白酶切割后产生 gp120 和 gp41，其氨基端与其他分子以非共价键形式结合。gp120 有 24 个糖基化位点，糖基化是 gp120 与 CD4 受体结合所必需的。已通过重组和突变实验定位了 gp120 蛋白的多个功能区，简述如下。

（1）主要抗原决定簇　包括 V3 区(第 301~324 位的环状肽段)的主抗原决定簇及若干个较弱的决定簇，其中有两个分别位于 gp41 的 616~632 位和 724~751 位。

（2）T 细胞决定簇　两个辅助性 T 细胞决定簇 T2 和 T1 分别在 105~117

表 9-5 HIV-Ⅰ、HIV-Ⅱ和 SIV 的辅助受体及表达模式分析

| 辅助受体 | 病　毒 | 细胞类型或组织 |
|---|---|---|
| CCRI | HIV-Ⅱ | 活化的体细胞,单核细胞,树突细胞 |
| CCR2b | HIV-Ⅰ,HCV-Ⅱ SIV | 单核细胞,T 细胞 |
| CCR3 | HIV-Ⅰ,H1V-Ⅱ | 嗜酸粒细胞,树突细胞,Th2 |
| CCR4 | HIV-Ⅱ | Th2,脑 |
| CCR5 | HIV-Ⅰ,HIV-Ⅱ,SIV | 活化的 T 细胞,单核细胞,树突细胞 |
| CCR8 | HIV-Ⅰ, HIV-Ⅱ, SIV | Th2,腺细胞,脑 |
| CCR9 | HIV-Ⅰ | 淋巴细胞,脑,胎盘 |
| C×CR2 | HIV-Ⅱ | 中性粒细胞,脑 |
| C×CR4 | HIV-Ⅰ,HIV-Ⅱ,SIV | 淋巴细胞,单核细胞,脑 |
| C×CR5 | HIV-Ⅱ | B 细胞 |
| ROC1 | HIV-Ⅰ,HIV-Ⅰ,SIV | 淋巴细胞,脑 |

表 9-6 HIV 编码的主要蛋白及其功能

| 基因产物 | 功　能 |
|---|---|
| Gag | 核心蛋白 |
| Pol | 多种酶功能 |
| Env | 包膜蛋白 |
| Tat(Tat-3、Ta) | RNA 沉默抑制因子,转录激活子 |
| Rev(Art、Trs) | 调控 RNA 剪切和运输,与 RRE 结合增加 *env* 的翻译 |
| Vif(Sor、A、P′、Q) | 侵染因子,病毒在巨噬细胞扩散中所必需 |
| Vpr(R) | 增加病毒复制 |
| Vpu | 辅助病毒组装和释放,以及 gp160/CD4 复合体的解聚 |
| Nef(3′ORF、B、E、F) | 形成同源二聚体,引起多效应性的影响,如负调控 CD4 |

位的 C1 区和 421~436 位的 C4 区，一个主要细胞毒 T 细胞决定簇位于第 308~322 位,与 V3 区有部分重叠。另有 3 个较弱的细胞毒 T 细胞决定簇在 gp41 上。

(3) CD4 受体结合区　该区位于 423~427 位(C4 区)。在上述功能区以外的某些氨基酸突变也能影响 gp120 的功能,说明上述各功能区发挥作用还依赖于整个分子特定的空间构象。

### 9.2.3 HIV 的复制

HIV 主要侵染人体的 T 淋巴细胞及巨噬细胞，并通过病毒膜蛋白 gp120 与细胞表面的 CD4 受体蛋白结合。CD4 分子含有两个免疫球蛋白结构域,它在人体内所结合的正常配体是Ⅱ型 MHC 抗原。HIV 与受体结合后,病毒核心蛋白和两条 RNA 链进入细胞(图 9-15)。反转录酶以病毒 RNA 为模板合成单链 DNA,并由宿主细胞 DNA 聚合酶合成双链 DNA(原病毒),经环化后进入细胞核并整合到染色体上,病毒核酸随细胞的分裂而传至子代细胞,可长期潜伏。主要过程如下:①原病毒整合到宿主染色体上,无症状;②原病毒利用宿主细胞的转录和合成系

统转录产生病毒 mRNA，其中一部分编码病毒蛋白，与基因组 RNA 组装成新的病毒颗粒，从宿主细胞中释放出来侵染其他健康细胞；③宿主细胞瓦解死亡。

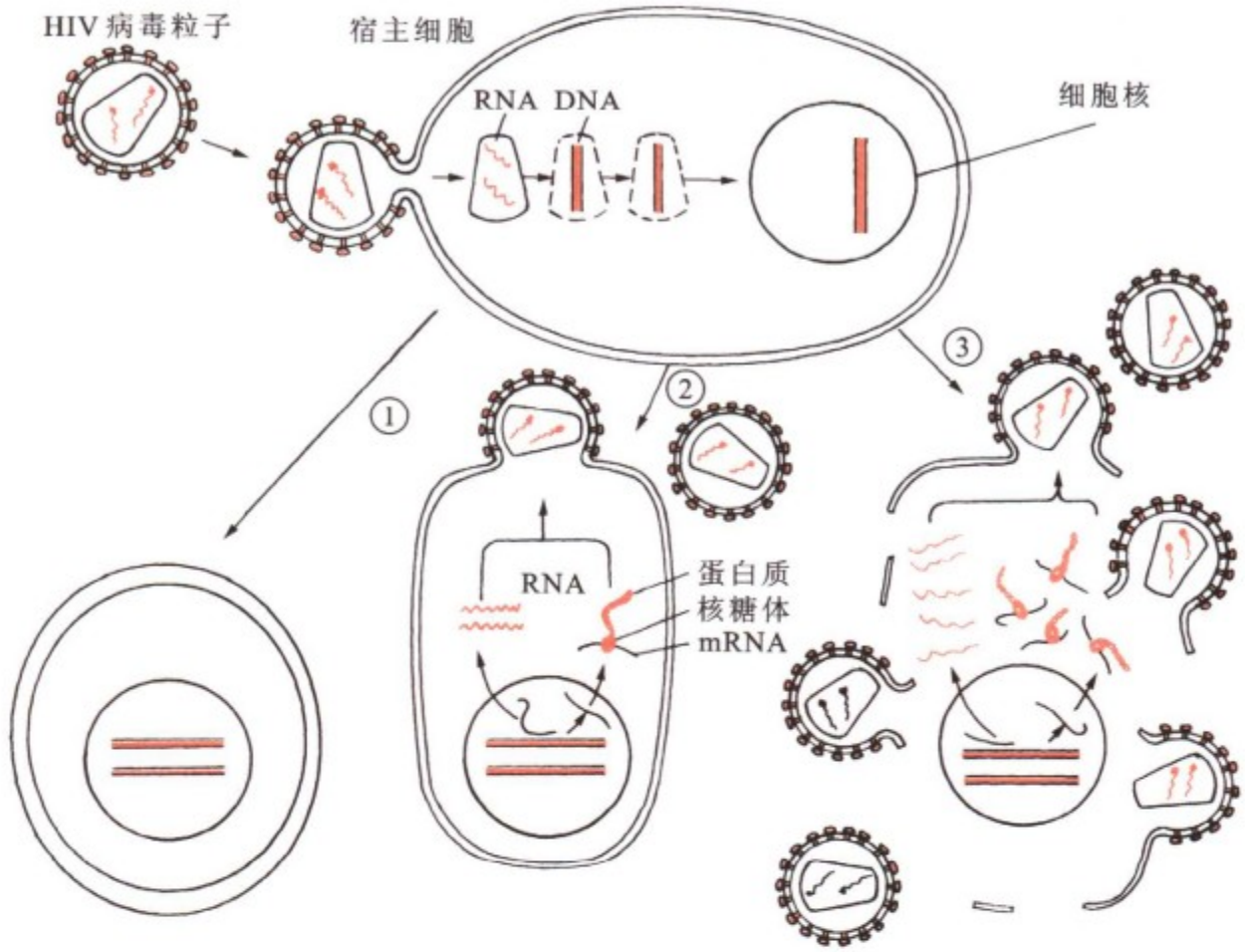



图 9-15 HIV-Ⅰ在人体细胞内的复制和侵染过程示意图

对 HIV-Ⅰ编码的 RNA 酶 H 和反转录酶的 X 线衍射晶体结构分析发现，合成前病毒第一条 DNA 链的过程中没有宿主细胞 DNA 酶的参与，完全由 HIV-Ⅰ反转录酶完成。该酶由两个 N 端完全一致的亚单位(p66 和 p51)组成，具有 DNA 聚合酶活性，其 N 端功能区以病毒 RNA 为模板催化合成互补 DNA 链。此外，p66 的 C 端功能区还能在 p51 亚基的协同作用下降解病毒模板 RNA 链。

HIV 的转录从基因组 5′端 LTR 起始，先形成一条长链 RNA 并迅速切割成短 mRNA 分子，用于指导合成 HIV 调控蛋白，但不翻译结构蛋白。当调控蛋白 Rev 达到其阈值后，转录由早期向晚期转变。此时，主要是转录长链 mRNA 并指导合成结构蛋白，组装成完整的病毒颗粒。HIV 复制过程的这种复杂调控机制使 HIV 感染细胞后产生不同的结果。在复制和表达量非常高的情况下可立即杀死细胞，低水平复制和表达时仅造成细胞功能失常，而不复制、不表达时病毒基因组以 DNA 形式与染色体整合，与宿主长期共存。这就是艾滋病在临床上的潜伏期长短不一、病程快慢不一的主要原因。HIV 前病毒整合是随机的，并不要求特定的 DNA 序列，因而可以整合到染色体上的任何部位。

### 9.2.4 HIV 基因表达调控

与大多数反转录病毒相比，HIV 的最大特点就是含有许多调控基因。这些

基因编码相应的调控蛋白，它们在病毒 RNA 的转录、转录后加工、蛋白质翻译、包装以及病毒颗粒的释放等各个过程中发挥重要作用。在这些过程中，除 HIV 自身的调控蛋白外，还有许多宿主细胞的调控蛋白参与。因此，HIV 复制的调控是一个十分复杂的系统。

1. LTR 序列

LTR 位于 *HIV* 基因组两端，其序列是高度保守的，在 HIV-Ⅰ与 HIV-Ⅱ及 SIV 之间的保守性高达 95%以上。LTR 5′端含有 *HIV* 基因调控所必需的多个特定调控区（真核类增强子和启动子单位），现已根据各区调控功能的异同分为调控单位、核心单位和反式激活效应元件单位，并在 LTR 中发现了许多细胞转录因子结合位点（图 9-16）。

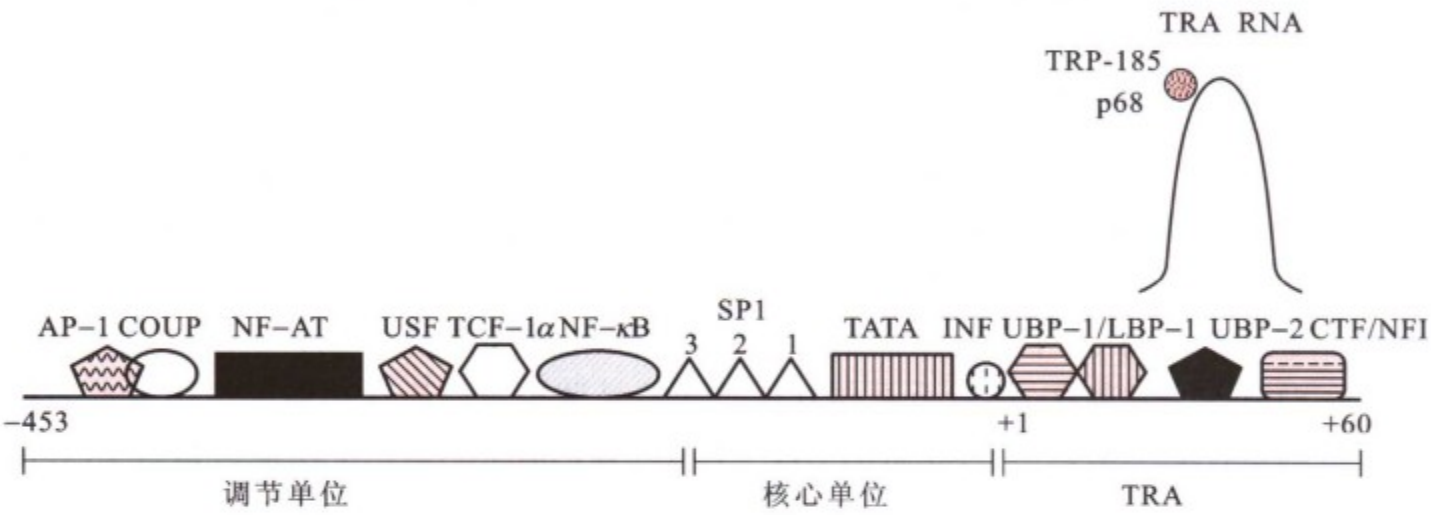



图 9-16 *HIV* 基因组 LTR 区中的 DNA 和 RNA 结合蛋白的作用位点

（1）核心调控元件　该元件从 LTR 起始延伸到-78 位核苷酸，至少包含 6 个可与细胞转录调控因子结合的区域，这些转录因子及其在 LTR 的结合部位分别是：鸡卵蛋白转录因子（COUP-TF），结合于 LTR-371~-334 位，激活蛋白 1（activator protein 1，AP1）结合于 LTR 的-347~-329 位。激活 T 细胞核因子（NFAT）在 LTR 有两个结合位点，分别在-292~-255 位和-216~-203 位。上游激活因子（USF）结合于-173~-160 位，T 细胞因子-1α（TCF-1α）结合于-139~-124 位，核因子 κB（NF-κB）结合于-104~-81 位。除 COUP-TF 和 USF 对 HIV 的转录起负调控作用外，其余均为正调控因子。在几种 DNA 结合蛋白中，AP1 是由含亮氨酸拉链结构的 C-Jun 和 C-Fos 家族成员组成的二聚体。COUP-TF 家族中的低相对分子质量蛋白质含有锌指结构，USF 为螺旋-环-螺旋结构，TCF-1α 含有高速泳动族模体（HMC motif），而 NF-κB 则含有锌指结构及锚定蛋白重复序列，该蛋白的结合位点也称为增强子区，由间隔 3~4 个核苷酸的两个 10bp 的 GGGACTTTCC 序列组成，负责调控 *HIV* 基因在多种细胞特别是 T-淋巴细胞中的高效表达。类似的序列在其他病毒如 SV40、巨细胞病毒及白细胞介素-2 受体、β-干扰素等基因的特异性和诱导型表达过程中都发挥了重要的调控作用。

（2）核心转录单位　HIV-Ⅰ LTR 核心转录单位的结构与其他病毒或细胞启动子的转录起始区相似，有 TATA 区和 SP1 结合位点。这些序列对转录的激

活具有重要作用,一旦缺失将使 LTR 失去启动子活性。

在核心转录单位-78~-46 位点的富含 GC 区有 3 个连续的 SP1 结合位点。SP1 蛋白含有 DNA 结合结构域和两个富含谷氨酰胺的结构域，可形成二聚或三聚体。当 3 分子 SP1 与 LTR 结合后,由于彼此间以及 SP1 与周围蛋白的相互作用改变了 DNA 的空间结构,诱导了依赖于 DNA 的蛋白激酶对 SP1 的磷酸化作用,激活 *HIV-Ⅰ* 基因转录。HIV-Ⅰ LTR 的-28~-24 位含有 TATA 元件,其两侧还有一个 CAXXTG 的同向重复序列，该重复序列可能是转录因子螺旋-环-螺旋 DNA 结合蛋白的识别位点。

(3) 反式激活因子应答元件　在 *HIV* 基因组+1~+60 位的核苷酸序列是反式激活因子(Tat)激活 HIV-Ⅰ转录所必需的,称为反式激活应答元件(TAR)。该序列存在于 *HIV* 基因组 RNA 和原病毒 DNA 上，是 HIV 复制所不可缺少的重要调控位点。

2. 参与 HIV 复制的调控蛋白

(1) Tat 蛋白　该蛋白是一个转录激活因子，在 HIV-Ⅰ、HIV-Ⅱ和 SIV 中高度保守,是病毒复制所必需的。HIV-Ⅰ的 *tat* 基因与 *env* 编码区有部分重叠,它含有两个外显子，编码了由 101 个氨基酸残基组成的 Tat 蛋白。第一个外显子所编码的 67 个氨基酸组成的多肽同样具有反式激活作用(图 9-17)。Tat 蛋白有 3 个功能区,第一个是与其他已知转录调控蛋白激活功能区结构基本一致的酸性氨基酸区,与 Tat 激活 *HIV-Ⅰ* 基因表达有关。第二个功能区富含 Cys,在其 16 个氨基酸残基中就有 7 个 Cys，可与二价金属离子结合并形成 Tat 二聚体。在这 7 个 Cys 中有 6 个是维持 Tat 活性所必需的,替代其中的任意一个均可使 Tat 失活却不影响 Tat 与金属离子的结合。该肽段能与相对分子质量为 $3.0\times10^4$ 的宿主细胞核蛋白相结合。

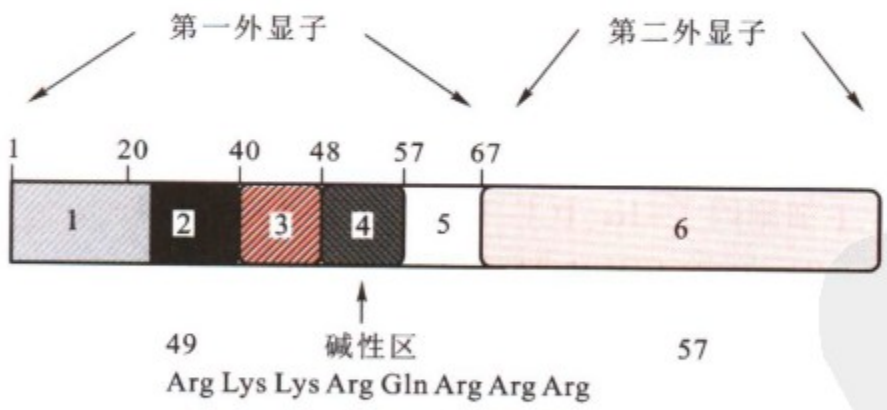



图 9-17　Tat 蛋白的结构示意图

Tat 的第三个功能区由一组带正电的碱性氨基酸残基构成，行使两个主要功能。一个是+48~+52 位的 GRKKR 序列,是核-核仁定位信号,可介导自身或异种蛋白进入细胞核。若将此肽段与 β-半乳糖苷酶融合即可使后者在核内聚集。该肽段第二个功能是与 TAR RNA 凸出部的特异性结合,通过 Tat 第 52 和第 53 位 Arg 残基与后者的磷酸基团相互作用得以实现。

(2) Tat 与 TAR 对 HIV 复制的协同调控作用　实验表明,Tat 和 TAR 结合并不能有效地激活 *HIV* 基因表达,还依赖于如下因素：

① TAR 序列及完整高级结构的影响　Tat 虽能结合环区有突变的 TAR 蛋白，但此时不能激活 *HIV* 的表达。

② 宿主细胞因子协同作用的影响　在不同细胞内，Tat 的活性可有上千倍的差异。在人鼠杂交细胞中，只有在保持人第 12 号染色体的杂交瘤细胞中，Tat 才能有效地激活 HIV-Ⅰ转录。研究发现，Tat 与 TBP Ⅰ结合形成复合物而发挥作用，说明 Tat 与 TAR 对 HIV 复制的调控还需宿主细胞编码蛋白的协同作用。

③ 对上游启动子和增强子的依赖　Tat-TAR 功能的发挥依赖于其上游的启动子及增强子，当 TAR 远离其上游启动子时，它的活性迅速下降。而且，这种依赖作用对非激活细胞和 Tat 激活细胞的表现是不一样的，增强子 SP1 序列在 Tat 诱导细胞中对 *HIV* 转录的促进作用远远大于非激活细胞，去除 SP1 序列甚至可提高非激活细胞中 *HIV* 的转录。而 NF-κB 序列在对两种细胞的作用正好与 SP1 相反。TATA 启动子序列的改变对未激活细胞影响很小，却可明显降低 HIV 在 Tat 诱导细胞中的复制。

研究认为，Tat 的主要作用可能是促进 RNA 聚合Ⅱ转录起始复合物的组装，因为在没有 Tat 蛋白存在的情况下，转录起始位点附近的顺式作用原件虽然也指导转录复合物的形成，但其延伸效率较低，只能转录出许多很短的产物。加入 Tat 蛋白后，短链 RNA 转录产物浓度迅速下降，取而代之的是长链 RNA。

最新的研究表明，Tat 蛋白是一个 RNA 沉默抑制因子。缺失突变和点突变结果表明，抑制 RNA 沉默功能区位于该蛋白的 38~72，且这段区域不具有转录激活功能，说明 Tat 蛋白的转录激活功能和 RNA 沉默抑制功能是分开的。*HIV-Ⅰ* 和 *Ⅱ* 也编码一个称为 Tax 的异源转录激活因子，它的主要作用是通过宿主细胞中的转录因子与 LTR 中的顺式作用元件结合，从而调控基因转录。

(3) Rev 蛋白　*rev* 基因由两个外显子组成，分别编码 25 个氨基酸和 91 个氨基酸的两个肽段，最终组装成有 116 个氨基酸、相对分子质量为 $1.9\times10^4$ 的调控结构蛋白表达活性的 Rev 蛋白(图 9-18)。它是一个重要的反式激活因子，调控 RNA 剪切和运输，与 RRE 结合增加 *env* 的翻译。它对 HIV 的许多调控蛋白的编码基因有负调控作用，而对其结构蛋白基因有正调控作用。Rev 蛋白的上述作用是通过与 *env* 和 *gag/pol* mRNA 上的一段 234 个核苷酸的 Rev 效应元件

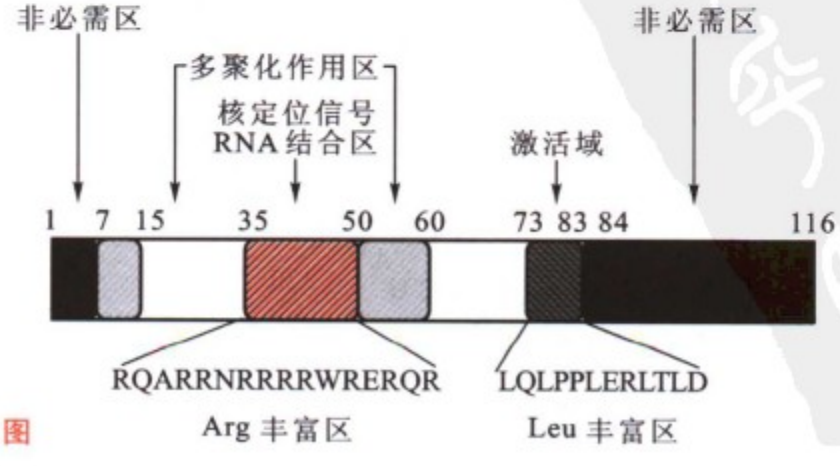



图 9-18　Rev 蛋白的结构示意图

实现的，是 *HIV*-Ⅰ特异的。它的主要功能是增加细胞质中未经编辑的基因组全长转录产物，增加含有 Rev 效应元件 RNA 的稳定性，并促进它们向细胞质转运。Rev 蛋白还通过阻碍剪接子的组装来抑制 RNA 的编辑，增加这些 mRNA 的翻译，从而促进 *HIV* 基因表达由早期（转录调控蛋白 mRNA）向晚期（转录 HIV 结构蛋白 mRNA）的转变。在 *rev* 基因缺陷型 HIV 前病毒中，仅有早期基因表达。只有加入 Rev 蛋白后，晚期基因才开始转录。Rev 蛋白还在结构蛋白 mRNA 进入细胞质的过程中发挥作用。

已经证实，Rev 蛋白和 RRE（Rev response element）的结合是通过 Rev 蛋白中富含碱性氨基酸的一段 14 肽（EGTRQARNRRRRWR）与 RRE 二级结构Ⅱ B 茎区特异性结合实现的（图 9-19）。第一个 Rev 蛋白与 RRE 的结合有助于其他 Rev 蛋白的结合，这些蛋白不是结合在 RRE 的特异位点上，而是直接结合到第一个 Rev 蛋白上，形成多聚体。Rev 蛋白功能的实现依赖于它向核内的转运，而这一过程是通过 Rev 蛋白的核定位信号功能区完成的。宿主细胞蛋白可结合到 Rev 蛋白近 C 端的功能区，协助加工和转运 *HIV* mRNA。

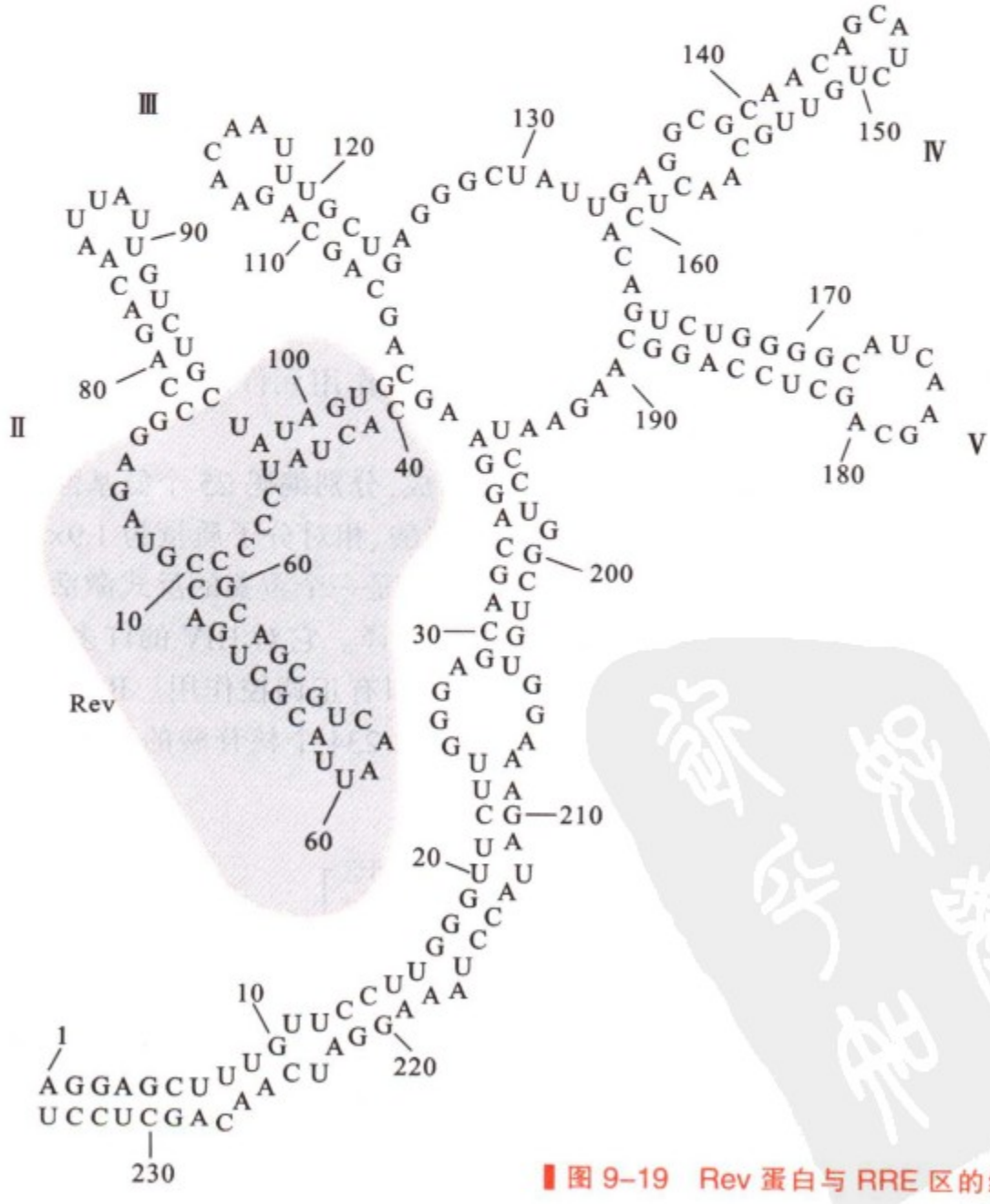



图 9-19 Rev 蛋白与 RRE 区的结合

（4）Nef 蛋白　一般认为，Nef 蛋白是一个负调控因子和磷酸化蛋白，相对分子质量为 $2.7\times10^4$，它不是 HIV 复制所必需的，但可抑制由 *HIV* LTR 特异性转录的 *HIV*-Ⅰ前病毒基因的表达。研究发现，烷基化后定位于细胞膜上的 Nef 蛋白比未烷基化并弥散于细胞核及细胞质内的 Nef 蛋白抑制 *HIV*-Ⅰ基因表达的效率比后者高 10 倍，推测 Nef 可能是通过膜细胞信使分子获得对 *HIV* LTR 的调控作用。Nef 的存在可提高和维持 HIV 在体内的含量，并在 HIV 病理过程中起作用。此外，Nef 蛋白形成同源二聚体，引起多效应性的影响，如负调控 CD4。

（5）Vpr 蛋白　该蛋白又称病毒 R 蛋白，由 96 个氨基酸组成，它不是 HIV 复制所必需的，但存在于病毒颗粒中。含有 *vpr* 基因的 HIV 毒株与无 *vpr* 基因的 HIV 毒株相比，前者在细胞中生长较快，引发细胞病变也较强。无 *vpr* 基因的 HIV-Ⅱ不能在巨噬细胞中繁殖。Vpr 蛋白可作用于 HIV-Ⅰ的 LTR 区，使它所调控的 HIV 复制速度提高 2~3 倍，说明 Vpr 在病毒早期侵染中有作用。

（6）Vpu 蛋白　这个蛋白是 HIV-Ⅰ特有的，其他灵长类动物免疫缺损病毒以及 HIV-Ⅱ和 SIV 均无此基因。它编码由 81 个氨基酸组成的磷酸化蛋白，称为病毒 U 蛋白，主要存在于细胞膜中。*Vpu* 基因的翻译受 Rev 调控，HIV 复制时并不需要 *vpu* 基因的表达，但若缺失 *vpu* 基因可使传染性病毒粒子的产量降低 5~10 倍，大大减少了病毒颗粒在细胞膜上的释放。据此，人们推测 Vpu 蛋白可能促进 HIV-Ⅰ病毒粒子的组装、成熟和释放。此外，Vpu 蛋白有助于 gp160/CD4 复合体的解聚。

（7）Vif 蛋白　该蛋白是相对分子质量为 $2.3\times10^4$ 的病毒传染性因子，是 HIV-Ⅰ侵染所必需的。缺失 *vif* 基因的 HIV-Ⅰ虽然仍在细胞内正常复制并产生病毒粒子，但这些粒子的侵染性不及野生 HIV-Ⅰ病毒粒子的 1/1 000。Vif 蛋白是病毒在巨噬细胞扩散所必需的。

*HIV* 基因组全长不到 10 kb，起调控作用序列不超过其因组总长度的 1/10，但是它的调控机制却非常复杂，参与调控作用的病毒自主编码的调控因子包括 Tat、Rev、Vif、Vpr、Vpu 及 Nef 等，还有宿主细胞调控因子参与。这些因子与因子之间，因子与顺式元件之间的相互作用保持了 *HIV* 基因表达的动态平衡。图 9-20 揭示了 Tat、Rev 和 Nef 对 *HIV* 基因表达的综合性影响。

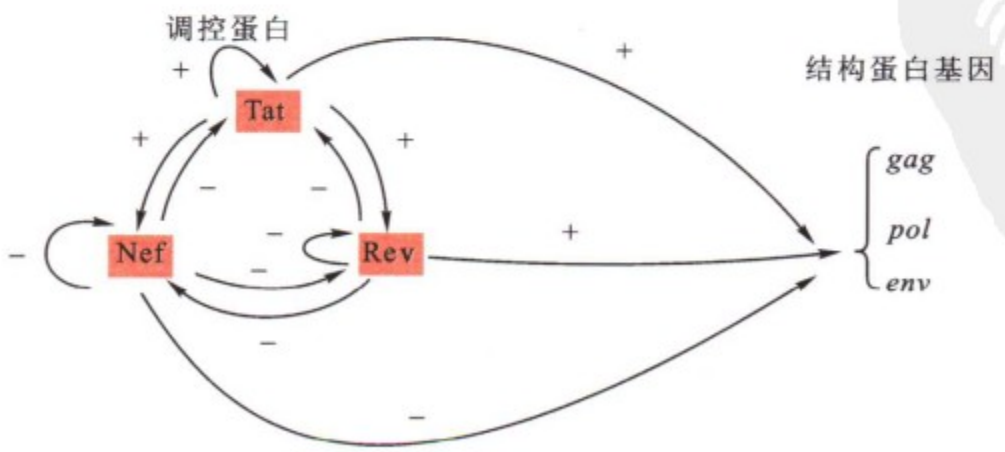



图 9-20　Tat、Rev 和 Nef 对 *HIV* 基因表达的影响分析

### 9.2.5 HIV 的感染及致病机制

HIV 初次感染人体后,立即大量复制和扩散,此时,感染者血清中出现 HIV 抗原,从外周血细胞、脑脊液和骨髓细胞中均能分离出 HIV。这是 HIV 原发感染的急性期。大约 70%以上的原发感染者在感染后 2~4 周内出现急性感染症状,包括发热、咽炎、淋巴结肿大、关节痛、中枢及外周神经系统病变、皮肤斑丘疹、黏膜溃疡等,持续 1~2 周后进入 HIV 感染的无症状潜伏期。在这个时期感染者无任何临床症状,外周血中 HIV 抗原含量很低甚至检测不到。但随感染时间的延长,HIV 重新开始大量复制并造成免疫系统损伤。临床上患者感染逐步发展到持续性全身性淋巴腺病(PGL)、艾滋病相关综合征(APC)等,直至发展到艾滋病。

在 HIV 自然感染过程中,机体可产生高滴度的抗 HIV 多种蛋白的抗体,主要是细胞类型特异性的中和抗体,其主要作用是在急性感染期降低血清中的病毒含量或部分清除病毒。HIV 的感染也刺激机体产生细胞免疫应答,对于杀伤 HIV 感染的细胞,阻止 HIV 经细胞接触而扩散等可能发挥重要作用,还能协助清除急性感染期 HIV 感染灶和 HIV 感染早期的少量有 HIV 复制的细胞。

尽管感染者具有对带有 HIV 的细胞和体液的免疫反应,但在大多数情况下机体的免疫力不足以清除 HIV,所以,一经感染 HIV 便终生携带病毒,随着时间延续,发病的可能性也逐渐增加。与一般病毒感染不同,HIV 感染的潜伏期可达数年,远远超过急性病毒感染,而发病率又远高于慢性和持续性病毒感染,感染后 7~8 年内已有约半数的感染者发病。产毒性 HIV 感染的最终结果是产生大量病毒颗粒,造成感染细胞死亡。这是反转录病毒科慢病毒亚科病毒的特征。HIV 除在细胞内大量繁殖造成细胞死亡外,还可通过以下几种途径导致免疫功能下降:①HIV 粒子表面的 gp120 蛋白脱落,与正常细胞膜上 CD4 受体结合,使该细胞被免疫系统误认为病毒感染细胞而遭杀灭;②因 T 细胞 CD4 受体被 gp120 封闭,影响了其免疫辅助功能;③HIV 的 gp120 蛋白可刺激机体产生抗 CD4 结合部位的特异性抗体,阻断 T 细胞功能;④带有病毒包膜蛋白的细胞可与其他细胞融合形成多核巨细胞而丧失功能。

研究还发现,HIV 除主要感染富含 CD4 蛋白的 T 细胞外,还可感染只带少量 CD4 蛋白或无此受体蛋白的细胞,包括 B 细胞、巨噬细胞、树突细胞、郎格汉斯细胞、肠上皮细胞、脑内的星形细胞和小神经胶质细胞以及毛细血管内皮细胞等。HIV 可侵染巨噬细胞并随该细胞游走至体内许多组织、细胞或脏器。HIV 还可侵染脑和肠细胞,因此,艾滋病患者的中枢神经系统常常发生病变,并伴有腹泻等症状。

尽管 HIV 有上述导致细胞病变的作用,但还不足以解释 HIV 感染所造成的免疫系统进行性损伤直至崩溃的严重病症。因为即使是艾滋病患者,其外周血淋巴细胞中也仅有极少数的细胞被 HIV 感染。最近有研究报道,鼠反转录病毒和鼠免疫缺损病毒感染的小鼠中病毒基因编码一种超抗原,可导致小鼠体内特定的 T 细胞亚群的消失和免疫缺损的发生,暗示艾滋病患者中也可能存在 HIV 编码的超级抗原,这可能是导致特定 T 细胞亚群缺失的原因。这些特定 T 细胞亚群的缺

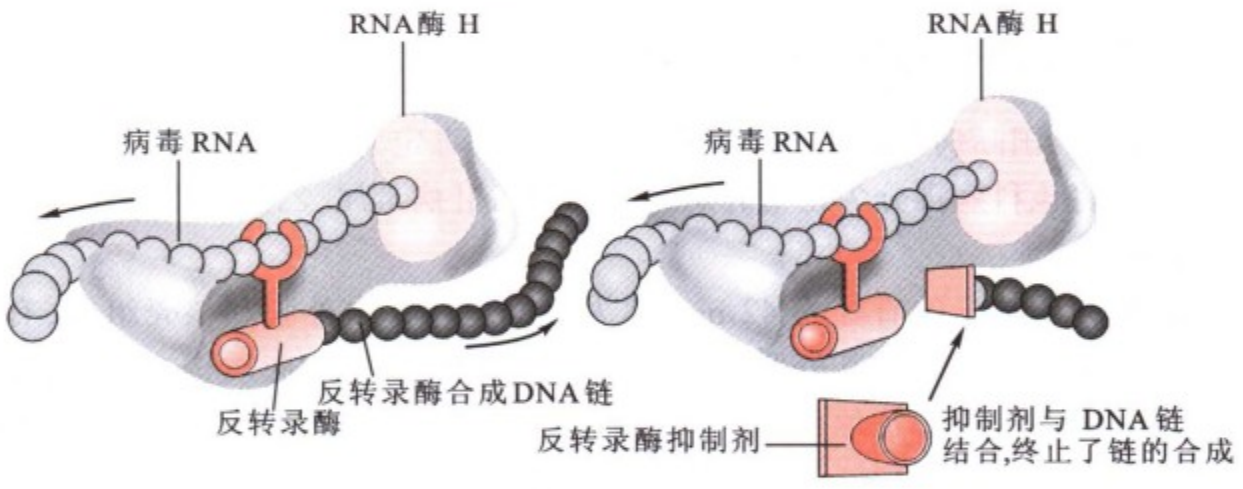
失造成机体的免疫防御网中出现漏洞，其结果使机体对某些致病微生物处于无反应状态。这种免疫病理机制可以解释艾滋病患者既出现了严重的免疫系统缺损，患者血液中却只有极少部分淋巴细胞带有 HIV 这种十分矛盾的现象。

### 9.2.6 艾滋病的治疗及预防

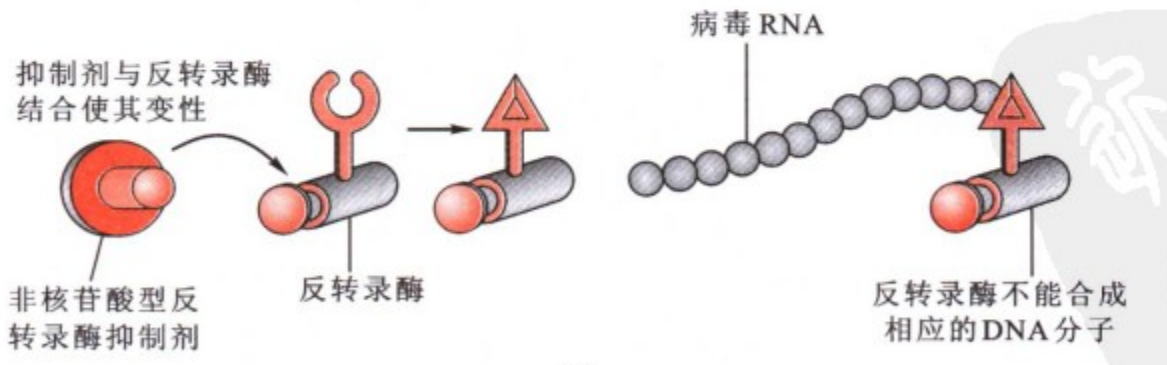
迄今为止，虽然对艾滋病的研究受到了高度重视，仍无任何药物可以完全抑制 HIV 在感染者体内的增殖并彻底治愈艾滋病。目前，已批准的抗 HIV 药物主要有以下几类（表 9-7）。图 9-21 是核苷酸型和非核苷酸型反转录酶抑制 HIV 的作用机制。

表 9-7 已被批准的抗 HIV 药物

| 核苷酸型反转录酶抑制剂 | 非核苷酸型反转录酶抑制剂 | 蛋白酶抑制剂 |
|---|---|---|
| Abacavir(ABC) | Efavirenz | Amprenavir |
| Stavudine(d4T) | Nevirapine | Indinavir |
| Didanosine(ddI) | Delavirdine | Lopinavir |
| Zalcitabine(ddc) | | Nelfinavir |
| Lamivudine(3Tc) | | Ritonavir |
| Zidovudine(AZT) | | Saquinavir |




(a)




(b)

图 9-21 核苷酸型和非核苷酸型反转录酶抑制剂的作用机制分析 (a)核苷酸型反转录酶抑制剂 由于它在结构上与脱氧核苷酸的相似性，掺入后使病毒 DNA 的合成不能进行；(b)非核苷酸型反转录酶抑制剂 与反转录酶相结合，通过限制该酶的移动性而影响它的活性。

除抗艾滋病药物的研制外，多种用于预防艾滋病的疫苗包括传统的减毒疫苗、灭活疫苗、重组病毒载体活疫苗、基因工程亚单位疫苗及合成寡肽疫苗等已经上市。由于 *HIV* 基因，特别是编码外膜蛋白的 *env* 基因的高度变异性，研制能有效抑制 HIV 病毒生长的疫苗困难很大。随着与 HIV 侵染有关的病毒基因产物及宿主相关受体分子三维结构与功能的逐步阐明，针对 HIV 不同生命阶段的特异疫苗会在不远的将来被研制出来，将对预防和控制艾滋病的发生发展具有重要意义。

## 9.3 乙型肝炎病毒——HBV

病毒性肝炎是严重威胁人类健康的世界性传染病，引起肝炎的病毒通称肝炎病毒(hepatitis virus)。目前已经知道的至少有甲肝病毒(HAV)、乙肝病毒(HBV)、丙肝病毒(HCV)、丁肝病毒(HDV)和戊肝病毒(HEV)这 5 种病毒。这些病毒在基因组结构、传播途径、临床症状和分类地位上很不相同。甲肝病毒属小 RNA 病毒科，乙肝病毒属嗜肝病毒科，丙肝病毒与黄热病毒和瘟病毒相似，为 RNA 病毒。丁肝病毒基因组由一个单链环状 RNA 分子组成，但其外壳能识别乙肝病毒的表面抗原，是一种与乙肝有关的缺陷型病毒，需要有 HBV 的辅助才能复制增殖。戊肝病毒基因组为单股正链有 poly A 的 RNA 病毒，球形无包膜，可能属于杯状病毒。

### 9.3.1 HBV 的分类及病毒粒子结构

1986 年，国际病毒命名委员会正式将乙肝病毒定为嗜肝 DNA 病毒科成员，1990 年又将该科病毒分为正嗜肝病毒属和禽嗜肝病毒属，乙肝病毒是正嗜肝病毒属成员。

HBV 完整粒子的直径为 42 nm，称为 Dane 颗粒，由外膜和核壳组成，有很强的传染性(图 9-22)。其外膜由病毒的表面抗原、多糖和脂质构成；核壳直径 27 nm，由病毒的核心抗原组成，并含有病毒的基因组 DNA、反转录酶和 DNA

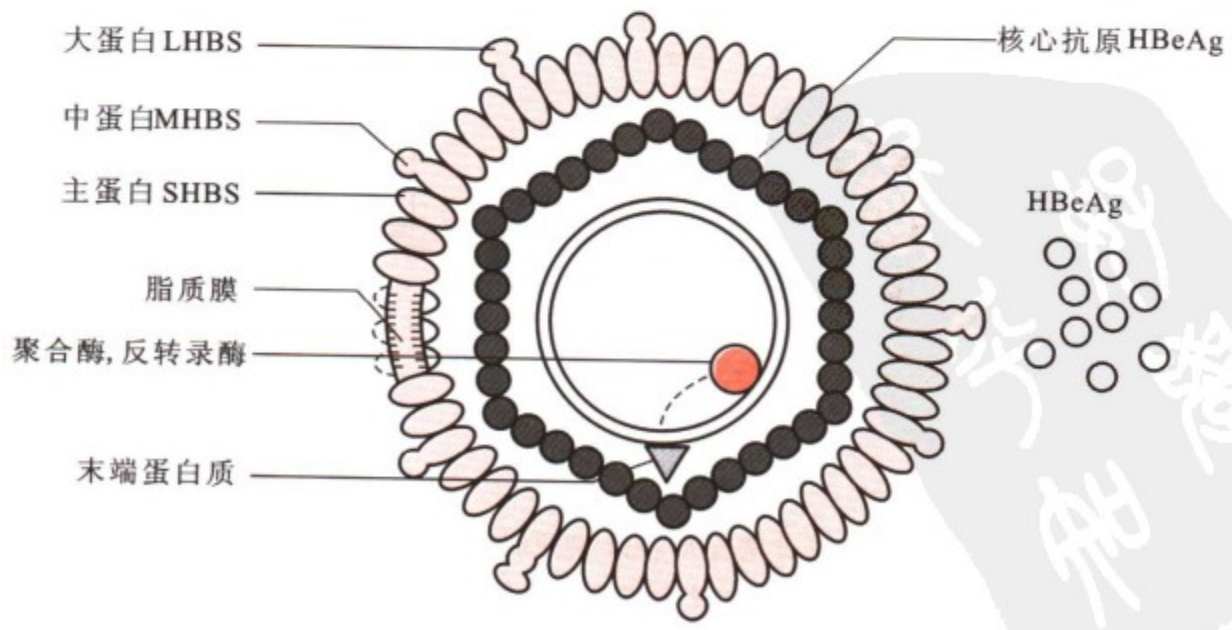



图 9-22 HBV 病毒粒子模型图

结合蛋白等。在慢性乙肝患者的血清中还有直径为 22 nm、长度不等的球状或棒状颗粒，它们只含表面抗原和脂质，不具有基因组 DNA，无侵染性。根据表面抗原性血清反应的不同，HBV 可以分为 *adr*、*adw*、*ayr* 和 *ayw* 等 4 个亚型。不同亚型的地区分布不同，我国大部分地区以 *adr* 为主，*adw* 次之；而西藏、新疆和内蒙等地则主要为 *ayw* 亚型。

## 9.3.2 HBV 基因组及其编码的蛋白

1. 基因组结构

乙肝病毒的基因组是一个有部分单链区的环状双链 DNA 分子，两条单链长度不一样。长链 L(3.2 kb)为负链，而短链 S 为正链，其长度不确定，约为负链的 50%~80%。因为它的 3′端随不同来源毒株而有不同的缺失，所以位置可变，产生了部分环状结构。基因组依靠正链 5′端约 240 bp 的黏性末端与负链缺口部位的互补维持了环状结构。长链的 5′端有一个共价结合的末端蛋白，可能是引物酶。而短链 5′端则通过共价键与具有帽子结构的短 RNA 结合。末端蛋白和短 RNA 都与病毒 DNA 的复制相关。此外，在两条链的互补区两侧各有一个 11 碱基的直接重复序列(5′TTCAC-CTCTGC-3′)，分别开始于第 1842 和 1590 核苷酸处，称为 DR1 和 DR2(图 9-23)。

自 1979 年第一次克隆和测定 HBV *ayw* 亚型全基因组序列以来，先后已有 20 多个不同的 HBV 基因组全序列被测定。分析发现，不同亚型中核苷酸水平的

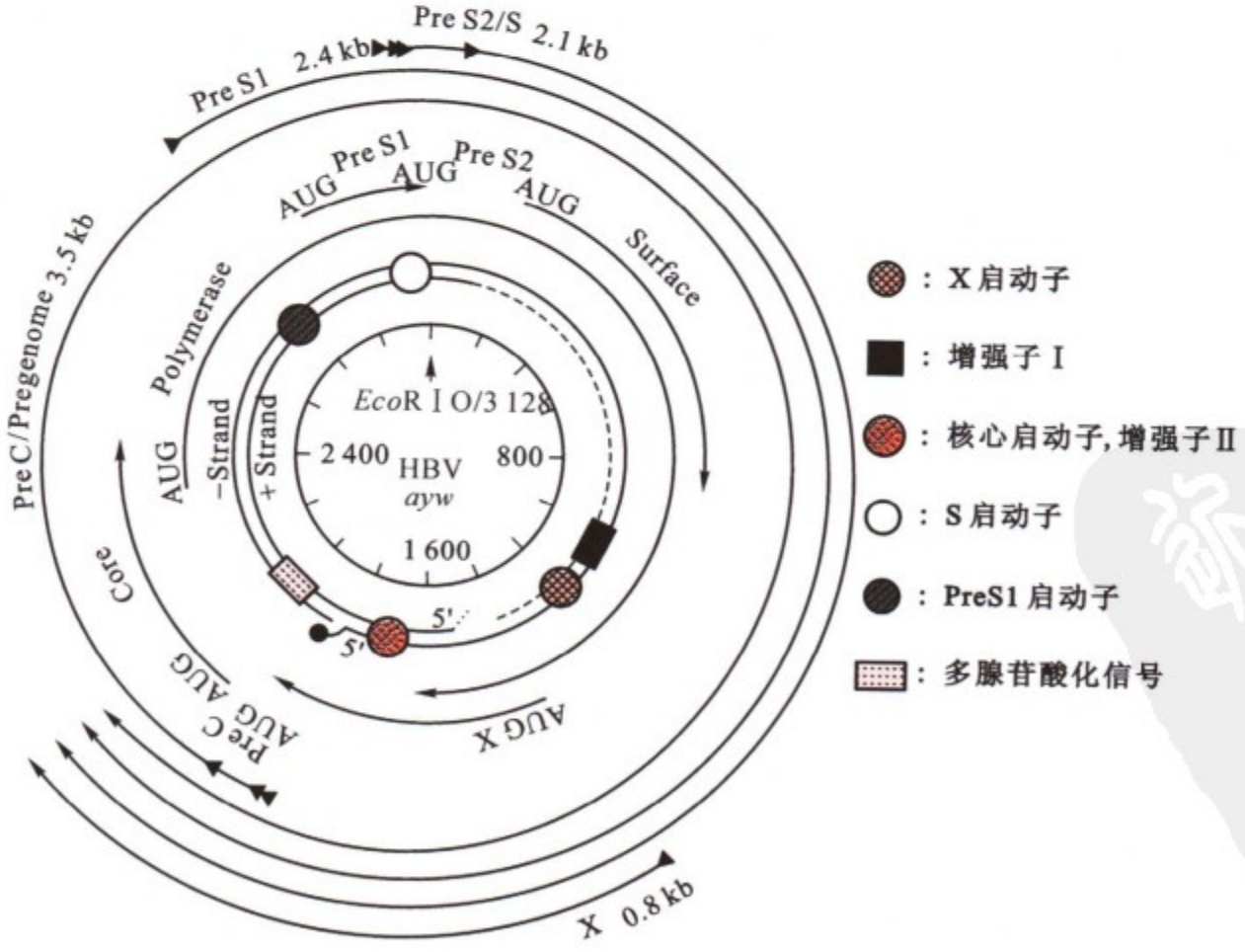



图 9-23 HBV 基因组结构(*ayw* 株)

变异约为10%左右。同亚型中的变异约为 2%左右,最高可达 11.5%。HBV 基因组的多态性是该病毒高变异率的分子基础，其生物学意义是近年来 HBV 研究中的热点之一。

2. HBV 转录产物

HBV 基因组结构的最大特点是功能单位非常密集,基因组高度压缩,编码区重复利用,所以有人也称为基因经济型基因组。它的核酸序列全部都是蛋白编码区。HBV 在感染过程中分别产生 3.5 kb、2.4 kb、2.1 kb、0.8 kb 4 种转录产物,都有 poly A 尾巴。其中 3.5 kb 和 2.1 kb 转录产物由 L 链录而来,2.4 kb 和 0.8 kb 转录产物由 S 链转录得到。3.5 kb mRNA 编码核心蛋白,其 3′端与 2.1 kb mRNA 的 3′端相同,比 HBV 基因组 L 链的全长还多 200 bp 以上,推测它的末端部分可能来自共价闭合的双链区。2.4 kb mRNA 编码 S 蛋白,2.1 kb mRNA 编码原 $S_1$ 和 $S_2$ 蛋白，而 0.8 kb mRNA 编码 X 蛋白。上述 4 种转录产物在病毒感染的细胞中相对含量有明显差异,其中 3.5 kb 和 2.1 kb mRNA 的转录量远远高于 2.4 kb 和 0.8 kb mRNA(图 9–23)。

3. HBV 基因转录的调控

(1) 顺式作用元件　HBV 的 4 种转录产物虽然有不同的 5′端，但都利用同一终止信号和 poly A 位点，它们的启动子序列已经被确定。C 启动子调控 3.5 kb mRNA 的转录起始过程,它定位于 1705~1805 位核苷酸处。在该启动子下游还发现一个具有启动子功能的区域,推测不同的 3.5 kb mRNA 可由不同的启动子控制。C 启动子上存在类似于 TATA 盒的序列。SP1 启动子位于–89~–77 核苷酸处,调控 2.4 kb mRNA 的转录起始过程。该启动子含有典型的 TATA 盒序列以及与该序列相距 45 个碱基的肝细胞特异性转录因子 HNF1 的结合位点。该位点主要存在于某些肝细胞专一表达基因的启动子区,HNF1 的结合是 SP1 启动子在分化的肝癌细胞株中高水平转录的必要条件。

SP2 启动子调控 2.1kb mRNA 的转录起始过程。它与 SV40 晚期启动子相似,位于 RNA 转录起始位点上游 200 个核苷酸内,可分为 A、B、C、D、E、F 和 G 7 个区,这些区都可能通过与特异的调控蛋白结合而影响转录水平。A、B、C 3 个区位于启动子远端(–168~–68 位核苷酸),若同时缺失,2.1 kb mRNA 转录水平下降到 30%以下。D 区位于–69~–49 处,是启动子的必需元件。近端的 E、F、G 3 三个区位于转录起始位点上游 45 个核苷酸内，该 DNA 序列与 SV40 主要晚期启动子有一定的序列相似性。F 为负调控区,E 区能抑制 F 区的负调控作用并与 G 区功能互补。B、D、E、F 4 个区可与相同或相似的转录因子结合,若将外源表达的转录因子 SP1 与病毒 DNA 进行共转染,可起始 HBV 基因转录。有实验发现,纯化的 SP1 可以直接与 B、D 和 F 区结合。

X 启动子位于–124~–24 核苷酸处,调控 0.8 kb mRNA 转录过程。有关该启动子的活性尚未见详细的研究报道。

(2) 增强子　HBV DNA 中存在两个可以激活 HBV 启动子转录的增强子区域。增强子Ⅰ位于表面抗原基因的 3′端,*X* 基因的 5′端,与 *X* 启动子相重叠。该增强子在肝细胞中对启动子的增强活性是非肝细胞的 10~20 倍,暗示它具有细胞特异性,这也可能是 HBV 嗜肝性的基础。

已在增强子Ⅰ序列中发现了多种细胞反式作用因子结合位点,NF-1a、NF-1b、NF-1c、AP1、C/EBP、EP 以及 X 蛋白等都可以与增强子Ⅰ相结合。根据主要的核蛋白结合位点,增强子Ⅰ被分为 2C、EP、E 和 NF-1 4 个区段。前两者为基础增强子,2C 与肝细胞专一性有关。若在基础增强子上加 E 和 NF-1 结合区,该启动子活性可提高 10 倍以上,其中 E 结合区可以与多种蛋白因子结合。E 结合区不仅能提高基础增强子的活性,还与 *X* 基因启动子的转录活性直接相关。一般认为,增强子Ⅰ能明显促进 SP1、SP2、X 和 C 启动子的转录。

增强子Ⅱ位于增强子Ⅰ下游 600 碱基处,是一个 148 碱基的 DNA 片段,根据其与主要核蛋白的结合位点分为 A、B 两个区。A 区为 60 碱基,是正调控元件,与增强子Ⅱ的肝细胞专一性有关。A 区单独存在时无增强子活性,必须与 B 区协同作用才有活性。B 区是增强子Ⅱ的基本单位,由 88 个碱基组成,单独存在时带有增强子Ⅱ活性的 60%~70%,又可分为 B1、B2 和 B3 三部分,其中前两部分是主要功能区,B2 是主要转录因子的结合位点。

(3) 反式作用因子　HBV 全序列被测定完成后,人们就发现了 *X* 基因,但其产物 X 蛋白具有反式调控因子功能则是近期才被证实的。研究发现,X 蛋白不仅能激活 HBV 自身启动子如 C、SP1、SP2 和 X 以及 HBV 增强子Ⅰ,还可以激活多种异源启动子和增强子,如 $\beta$-干扰素、HIV-Ⅰ、SV40 早期启动子以及一些细胞因子如Ⅰ型 MHC 等的表达。

X 蛋白虽然具有转录水平上的反式激活作用,但尚未发现它有 DNA 结合能力,说明 X 蛋白必须通过其他蛋白因子起作用。分析 X 蛋白对 HBV 增强子Ⅰ的作用时发现,该增强子的 E 元件十分重要,所以又将这个由 26 碱基组成的元件命名为 X 应答元件(X-responsive elememt,XRE)。X 蛋白与相关细胞因子结合后,再结合于靶序列上发挥作用。XRE 有 NF-κB、AP1、AP2 及 CREB 类似物的结合序列。近来有报导认为 X 蛋白本身具有丝氨酸/苏氨酸激酶活性。

4. HBV 的编码区及产物

(1) S 编码区　S 编码区编码乙肝表面抗原蛋白,分别编码由 226 个氨基酸残基的表面抗原主蛋白(SHBS)、108~115 个氨基酸的原 S1 蛋白和 55 个氨基酸组成的原 S2 蛋白 3 部分组成。S 蛋白和原 S2 蛋白组合在一起被称为中蛋白(MHBS),S 蛋白和原 S1、原 S2 蛋白组合在一起时被称为大蛋白(LHBS)。SHBS 是病毒外壳蛋白和 22 nm 颗粒表面抗原的主要成分,占病毒蛋白的 70%~90%,中蛋白和大蛋白则暴露于病毒颗粒表面,这 3 种蛋白可以部分或完全被糖基化。表面抗原有很强的免疫原性,是乙肝疫苗的主要成分。

(2) P 编码区　该区长 2532 碱基,约占全基因组 3/4 以上,是最长的编码区,包含全部 S 编码区并与 C 和 X 编码区有部分重叠。P 编码区由 3 个功能区和一个间隔区构成,排列顺序为末端蛋白(又称引物酶)、间隔区、反转录酶/DNA 聚合酶及 RNaseH。末端蛋白是病毒进行反转录时的引物。P 编码区可能先翻译成 $9.5\times10^5$ 多肽,然后加工成较小的功能型多肽。

(3) C 编码区　该区长 639 碱基,翻译产物为病毒核心抗原(HBcAg)。其原始翻译产物是前核心蛋白,切除 N 端 19 肽和富含精氨酸的 C 端后,成为 E 抗

原(HBeAg),相对分子质量为 $2.2\times10^4$ 蛋白,是核衣壳上唯一的结构蛋白。HBeAg 可分泌到 Dane 颗粒以外,存在于血清中。

(4) X 编码区　该区是最小的编码区,编码 X 蛋白,覆盖了负链的缺口部位,虽然长度不等,但主要产物由 154 个氨基酸残基组成。X 蛋白具有反式调控作用,能激活多个同源或异源启动子或增强子,与肝癌的发生有相关性。

### 9.3.3 HBV 的复制

乙肝病毒基因组虽为双链 DNA,但其复制并不通过半保留复制方式,而是通过如图 9-24 所示的反转录途径。病毒感染宿主后,首先脱掉外壳进入细胞质,基因组 DNA 进入细胞核,经 DNA 聚合酶修复正链缺失部分,形成共价闭环状双链 DNA 作为转录模板。此后,该环状 DNA 在感染细胞核中大量扩增,成为

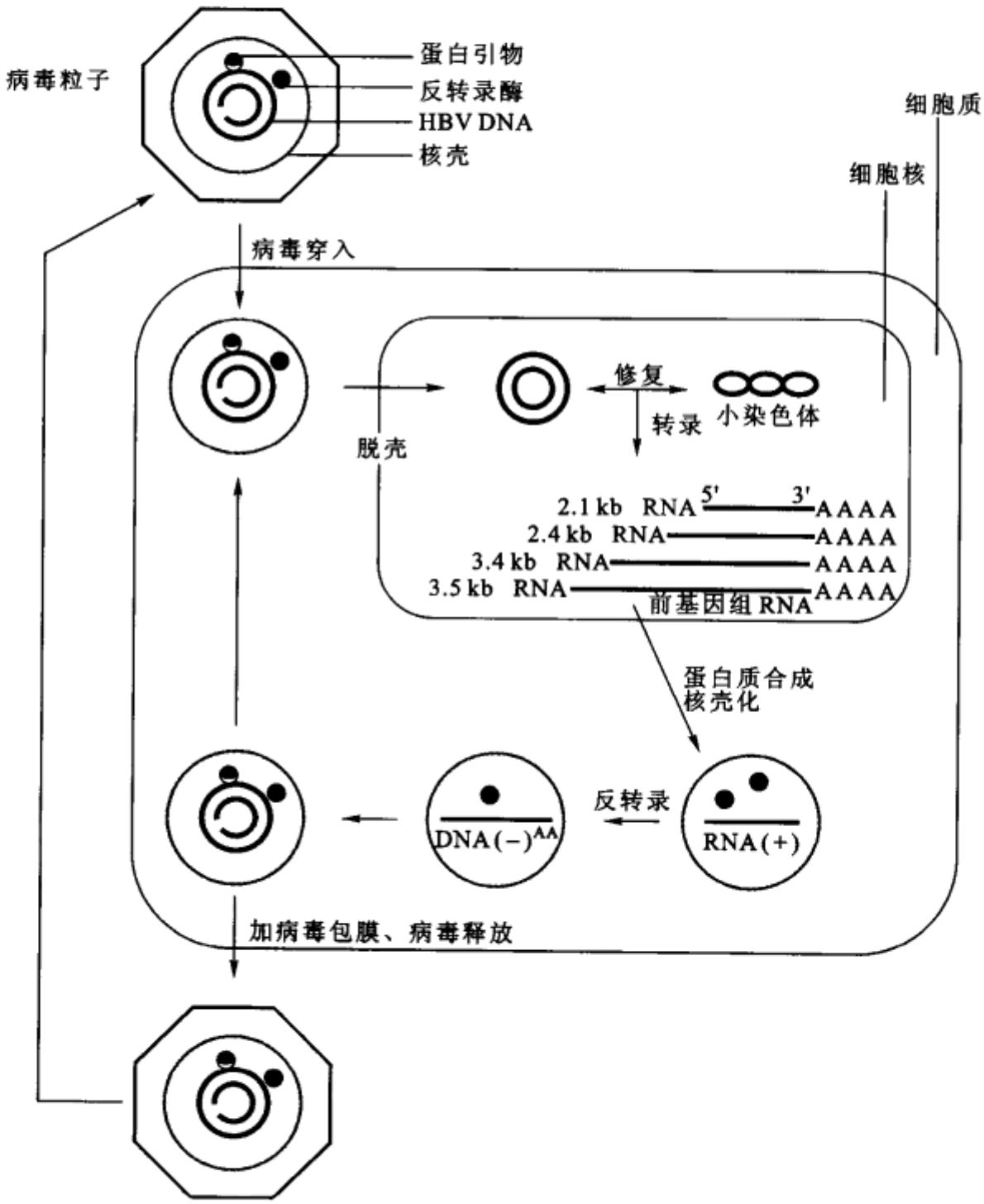



图 9-24　HBV 基因组复制模型图

HBV 基因组复制时的独特现象。

其次,HBV 基因组在宿主 RNA 聚合酶作用下开始转录，产生各种 mRNA，在这些 RNA 中，全长 3.5 kb 产物为前基因组 RNA，其 5′端自 DR1 处开始,自 5′向 3′延伸，经过 DR2 后继续合成至 DR1，最后终止于 DR1 后 85 碱基处的 TATAAA 序列,全长比病毒 DNA 多 130~270 个碱基。

第三步,前基因组中的部分 RNA 被包装到核心颗粒中并进行反转录。以与母本负链 DNA 5′端相连接的末端蛋白为引物,在反转录酶作用下,自 DR1 处开始反转录合成负链 DNA(图 9-25)。cDNA 合成后,由 RNaseH 将模板(前基因组 RNA)降解,但 5′端从帽子结构到 DR1 不被降解区域作为正链合成时的引物。

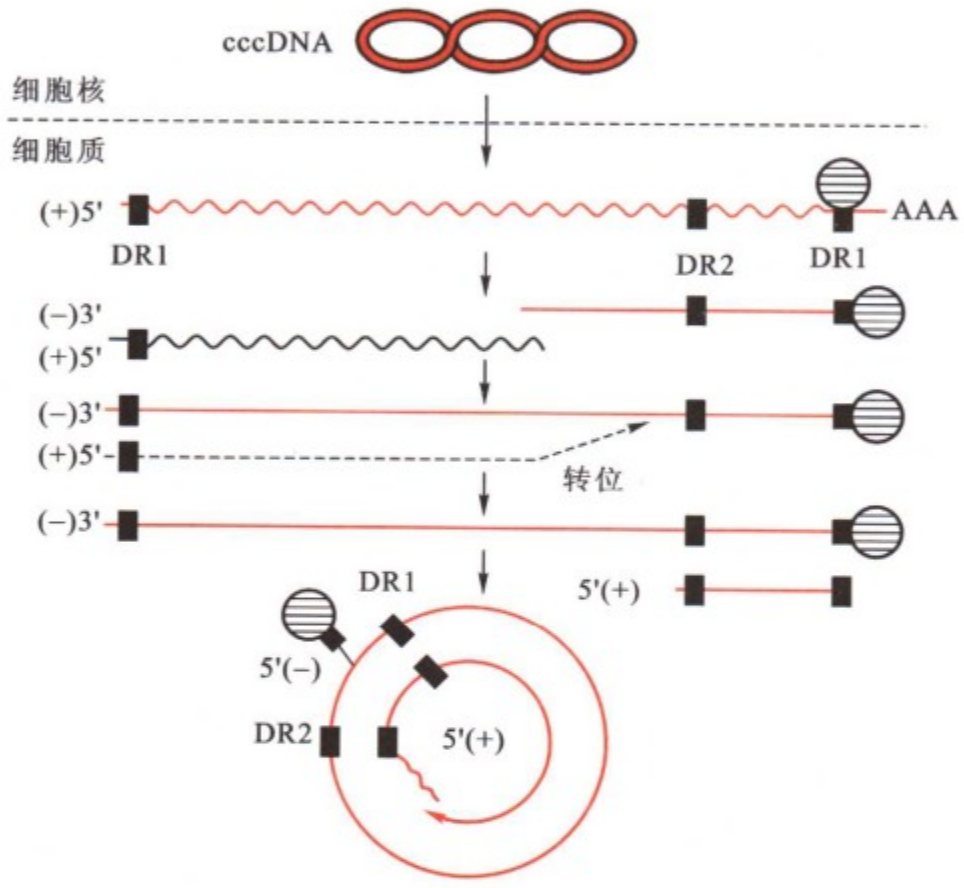



图 9-25 HBV 基因组的复制过程

第四步,在 DNA 聚合酶作用下合成正链,合成自 DR1 处开始,再经跳跃传位至 DR2,这一过程称为“引物易位”。因此,正链 5′端覆盖了负链的缺口,维持了双链 DNA 的环状。在某些情况下也可能不发生引物易位现象,导致线性双链 DNA 的合成。正链合成尚未完成时,就可能由于某些障碍而被终止。外壳蛋白便将核衣壳包装成病毒颗粒,形成了不完整的双链基因组。由于病毒复制增殖过程中部分酶类与细胞因子可能来自宿主细胞,所以,病毒的复制增殖过程也是干扰细胞正常代谢的致病过程。

## 9.4 人禽流感的分子生物学机制

人禽流感(human avian influenza,HAI)是人感染 AIV(avian influenza virus)

后引起的以呼吸道症状为主的临床综合征。引起 HAI 的病原为禽流感病毒，属正黏病毒科流感病毒 A 型流感病毒。禽流行性感冒(avian influenza，AI) 简称禽流感，又名真鸡瘟(fowl plague)、欧洲鸡瘟、鸡疫等，是由 A 型流感病毒引起的从呼吸系统病变到全身败血症的一种高度急性传染病。鸡、火鸡、鸭等家禽及野生禽类均可感染。美洲、欧洲、亚洲的许多国家和地区都曾发生过此病，对养禽业造成了巨大的损失。禽流感病毒不仅是人流感病毒株形成的最庞大的基因库，它还能直接感染使人发病，因此，禽流感已成为人类公共卫生事业面临的重大危害之一。

### 9.4.1 人禽流感病毒特点及分型

禽流感属正黏病毒科(*Orthomyxoviridae*) 流感病毒属的 A 型流感病毒。正黏病毒分为三个属，即 AB 型流感病毒属，C 型流感病毒属和类托高土病毒属(Thogoto like viruses)，病毒颗粒直径为 80~120 nm，平均 100 nm，典型的病毒粒子呈球状，基因组为多个负链 RNA 片段组成(图 9-26)。AIV 属 A 型流感病毒，是正黏病毒属中唯一感染禽类的病毒。禽流感病毒可按病毒粒子表面对红细胞有凝集性的血凝素(hemagglutinin，HA)和能将吸附在细胞表面的病毒子解脱下来的神经氨酸酶(neuraminidase，NA)糖蛋白进行分类，共有 16 个 H 亚型和 9 个 N 亚型。根据 HA 与 NA 的不同，又可将禽流感分为上百个血清亚型。导致禽流感暴发的主要是以 H5、H7 为代表的高致病性禽流感(HPAI)和以 H9 亚型为代表的低致病性禽流感(LPAI)毒株。

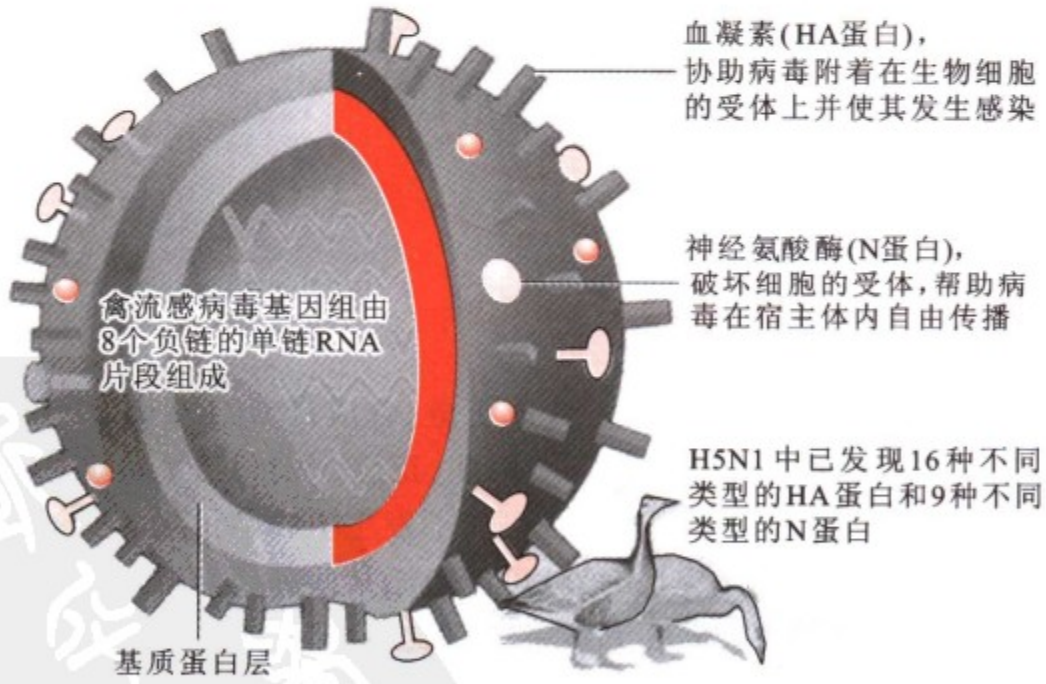



图 9-26 禽流感病毒颗粒主要结构蛋白及功能分析

1997 年，我国香港首次分离到由禽病毒跨越物种屏障同时感染鸡和人的 A 型流感病毒-H5N1。由于该病毒所有基因片段与禽流感病毒在序列上有很高的相似性，一般认为该病毒起源于禽类。当流感病毒开始感染新的宿主时，为了适

应新的环境和宿主细胞，会在强大的选择压力下进行病毒重组以形成新的株系或致病型。

## 9.4.2 禽流感病毒进入细胞及转录与复制

H5N1 禽流感病毒通过 HA 蛋白上的凝集素位点与细胞表面含唾液酸的糖蛋白受体结合，通过受体介导的细胞内吞作用进入细胞(图 9-27)。由于核内体(endosome)pH 较低，导致 HA 构象变化，细胞膜与病毒外膜发生融合，释放病毒衣壳进入细胞质。流感病毒启动转录时，病毒核酸内切酶将宿主细胞 mRNA 5′端的帽子结构切下作为病毒 RNA 聚合酶的引物，转录产生 6 个单顺反子的 mRNA，并翻译成 HA、NA、NP 和三种聚合酶(PB1、PB2 和 PA)。*NS* 和 *M* 基因的 mRNA 进行拼接后，每一个产生了两个 mRNA，依不同可译框架(ORF)进行翻译，产生 NS1、NS2、M1 和 M2 蛋白(表 9-8)。NA 是病毒核衣壳的主要成分，MA 是主要结构蛋白，存在于病毒囊膜内侧，分为 M1 和 M2。PB1，PB2 和 PA 是病毒复制酶的三个主要组分。NS1 和 NS2 为非结构蛋白，在病毒复制早期起重要作用，NS1 蛋白调节病毒 RNA 的合成、Pre-mRNA 的运输、剪切及 mRNA 的翻译

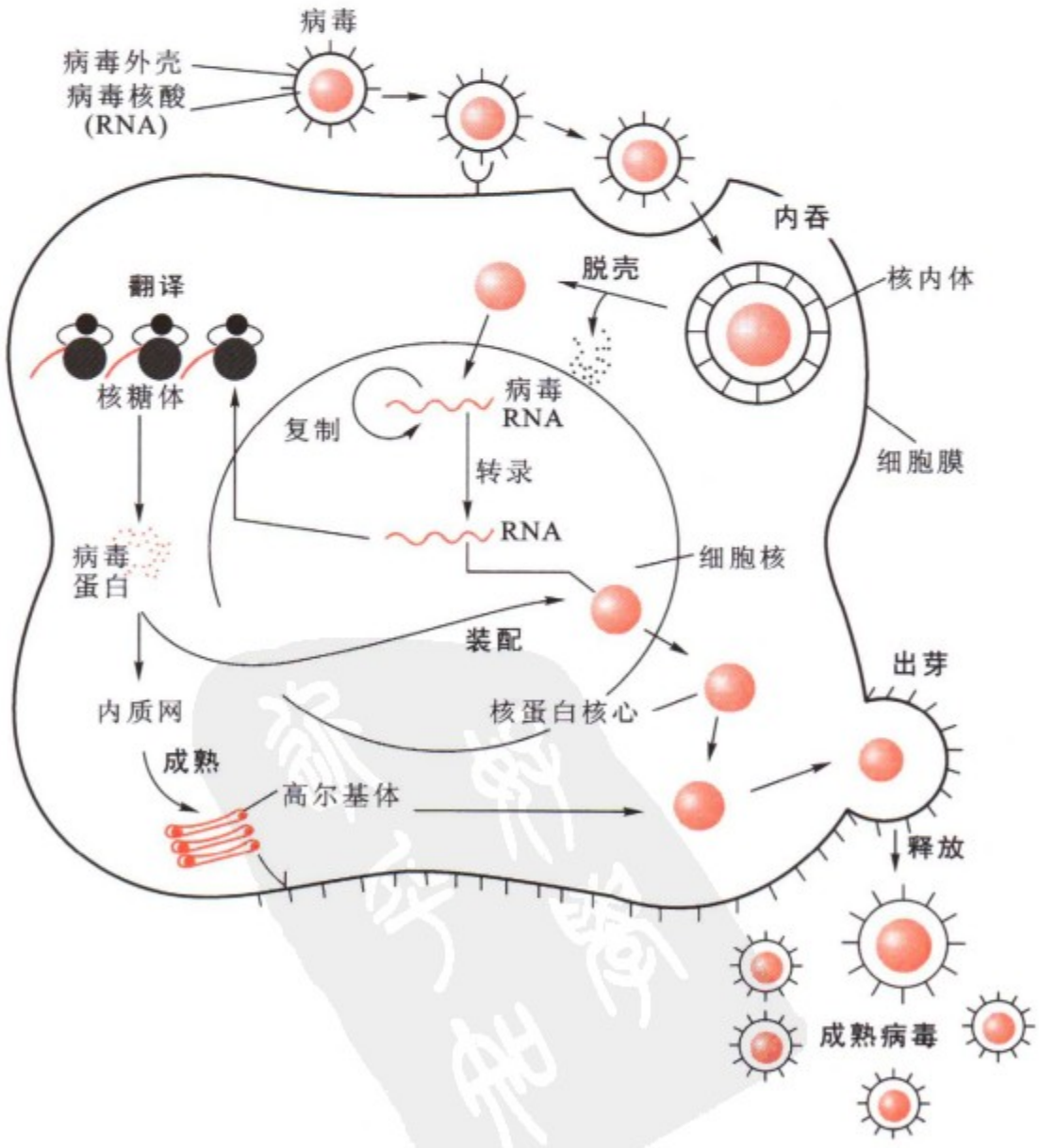



图 9-27 流感病毒 A 株进入宿主细胞及其复制过程图示

过程。NS2 蛋白与 M1 蛋白相互作用,参与病毒复制周期的调控。

表 9-8 禽流感病毒 RNA 在不同类型病毒中可能编码不同的蛋白质

| 流感病毒编码蛋白 | | |
|---|---|---|
| A 型 | B 型 | C 型 |
| PB2 | PB1 | P1 |
| PB1 | PB2 | P2 |
| PA | PA | P3 |
| HA | HA | HE |
| NP | NP | NP |
| NA | NA,NB | |
| M1,M2 | M1,M2 | M |
| NS1,NS2 | Ns1,Ns2 | Ns1,Ns2 |

### 9.4.3 禽流感病毒感染人类的机制

有关禽流感导致人类流感大流行的机制,目前主要有猪作为中间宿主的混合器假说和禽流感病毒直接感染人(即人本身充当混合器)两种假说。研究发现,禽流感病毒跨越种属间的屏障从家禽直接感染人与其基因的分子结构的特殊性有关,该病毒的基因组为分多个片段,极易发生基因的交换、重组并发生变异,造成感染宿主和致病性的改变。

1. HA 受体结合位点的突变导致禽流感病毒易于感染人类细胞

HA 是流感病毒表面主要的糖蛋白,在病毒感染过程中起关键作用,因为病毒借助 HA 对宿主细胞表面受体的识别作用而吸附于被感染细胞的表面。禽流感病毒与宿主细胞结合的特异性与 HA 结合受体位点的第 226 位氨基酸密切相关,如果该位点氨基酸残基为谷氨酰胺,则表现为 SAa22,3 Gal 受体(禽类呼吸道上皮细胞表面受体)结合特性;但若该位点为亮氨酸,则表现为 SAa22,6 Gal 受体(人呼吸道上皮细胞表面受体)结合特性;若该位点为甲硫氨酸,对 SAa22,3 Gal 和 SAa22,6 Gal 具有相同的结合能力。1997 年香港禽流感 H5N1 病毒就是由于第 226 位氨基酸由谷氨酰胺突变为甲硫氨酸,使其 HA 与受体结合的位点发生变异,导致流感病毒的宿主特异性改变,使禽流感病毒直接感染人细胞。对 H5 及 H7 亚型高致病力毒株的核酸序列和氨基酸序列分析表明,高致病力毒株的 HA 在裂解位点附近有多个碱性氨基酸残基,易被宿主蛋白酶系统裂解为 HA1 和 HA2,致病性明显增强。

2. PB2 蛋白 627 位氨基酸的点突变导致人禽流感病毒复制能力增强

禽流感病毒感染人体细胞后必须进行有效复制,以达到具有一定数量的群体,才能诱发机体的非特异性免疫反应,产生对机体的直接或间接的损伤作用。而禽流感病毒是否能在人体细胞内进行有效复制,主要与病毒复制酶(PB1、PB2 和 PA)基因有关。有研究发现,当 PB2 蛋白基因来源于人流感病毒,无论其

他基因片段来源于禽流感或人流感病毒,重组病毒在哺乳动物体细胞中都能有效复制;当PB2蛋白基因来源于AIV时,重组病毒则不能在哺乳动物细胞中进行有效复制。禽流感病毒株PB2的627位氨基酸都是谷氨酸,而人流感病毒该位点上是赖氨酸。香港H5N1毒株中PB2的第627位氨基酸就发生了由谷氨酸型向赖氨酸型的转化,因此,有人认为,PB2的627位氨基酸是决定A型流感病毒宿主范围的关键位点。

3. NS1第92位氨基酸的突变导致人禽流感病毒致病能力显著增强

人禽流感H5N1病毒感染后与普通的人流感病毒相比具有极强的致病性。研究表明,人禽流感H5N1亚型病毒的高致病性与其非结构蛋白有密切的关系。普通流感病毒感染时,非结构蛋白(NS1)产生于机体受感染的细胞,与病毒复制产生的双链RNA相互作用,产生的细胞内信号作用于RNA病毒蛋白激酶(PKR)以及核因子NF2KB,启动干扰素基因的转录和干扰素的合成,从而发挥抗病毒的非特异性免疫作用。在高致病性的人禽流感H5N1亚型病毒感染者体内虽然存在高浓度的干扰素和肿瘤坏死因子2α,这些细胞因子却不能抑制或杀死高致病性的人禽流感病毒,因为这种病毒不仅对干扰素和肿瘤坏死因子2α具有极强的抵抗力,而且在细胞内增殖速度极快,加速了炎性细胞因子的产生,引起全身性炎症反应综合征,从而表现出极高的致病力。造成这种现象的原因可能与H5N1亚型病毒的非结构蛋白基因(NS1第92位氨基酸)发生变异有关。高致病性的人禽流感H5N1病毒非结构蛋白能诱导机体淋巴细胞凋亡,显著降低CD4/CD8的比值,使机体的免疫功能下降,可能也是人禽流感H5N1病毒具有高致病性的原因之一。但是,目前尚不清楚非结构蛋白变异后,病毒通过何种信号途径逃避细胞因子抗病毒作用。

## 9.5 严重急性呼吸系统综合征-SARS的分子机制

### 9.5.1 SARS-CoV的结构与分类

严重急性呼吸系统综合征(severe acute respiratory syndrome,SARS)已经证明是由SARS冠状病毒引起(SARS-CoV)。SARS-CoV属冠状病毒科(*coronaviridase*)、冠状病毒属(coronavirus),一般呈多形态,病毒颗粒直径在80~160 nm,有囊膜,囊膜表面镶嵌有12~24 nm的球形梨状或花瓣状纤突(Spike),纤突之间有间隙。由于囊膜上的纤突规则排列呈皇冠状,故称为冠状病毒(图9-28)。病毒颗粒的囊膜由两层脂质组成,在两层脂质中镶嵌有两种糖蛋白,纤突蛋白S(又称E2)和血凝素酯酶HE(hemagglutinin estarase,又称E3),还有膜蛋白M(又称E1)和小包膜糖蛋白E(envelope pretein)。病毒粒子内部为核衣壳蛋白N和核基因组RNA组成的核蛋白核心。

已知冠状病毒分为3组,第1和第2组为哺乳动物病毒,第3组是禽类病毒。基因序列分析表明,SARS-CoV的基因组序列与其他三组都存在很大的差异,很难被归入某一组(图9-29)。虽然有学者认为它可能是第2组的重组病毒,

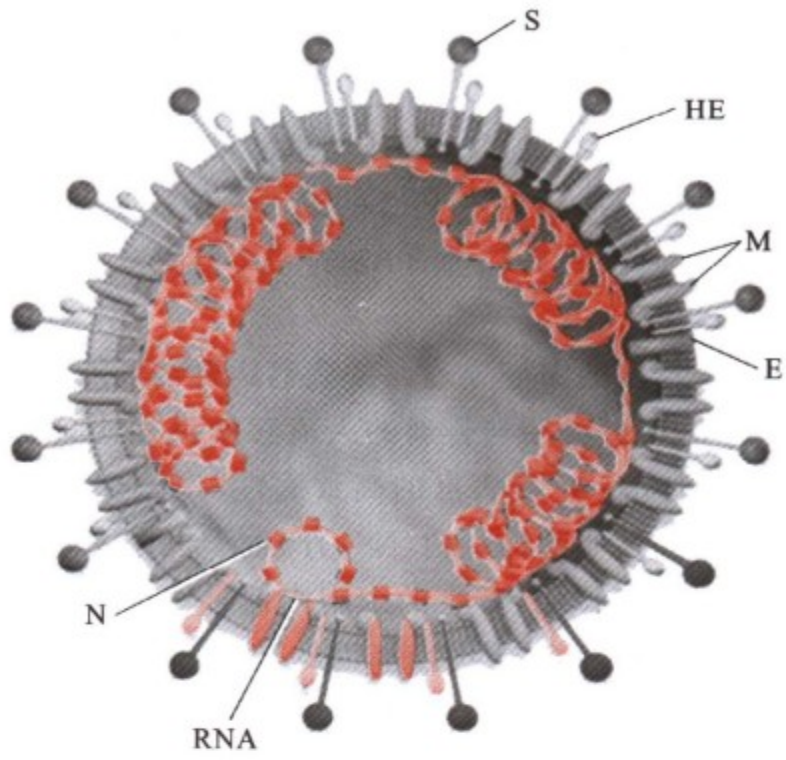



图 9-28 冠状病毒颗粒结构示意图

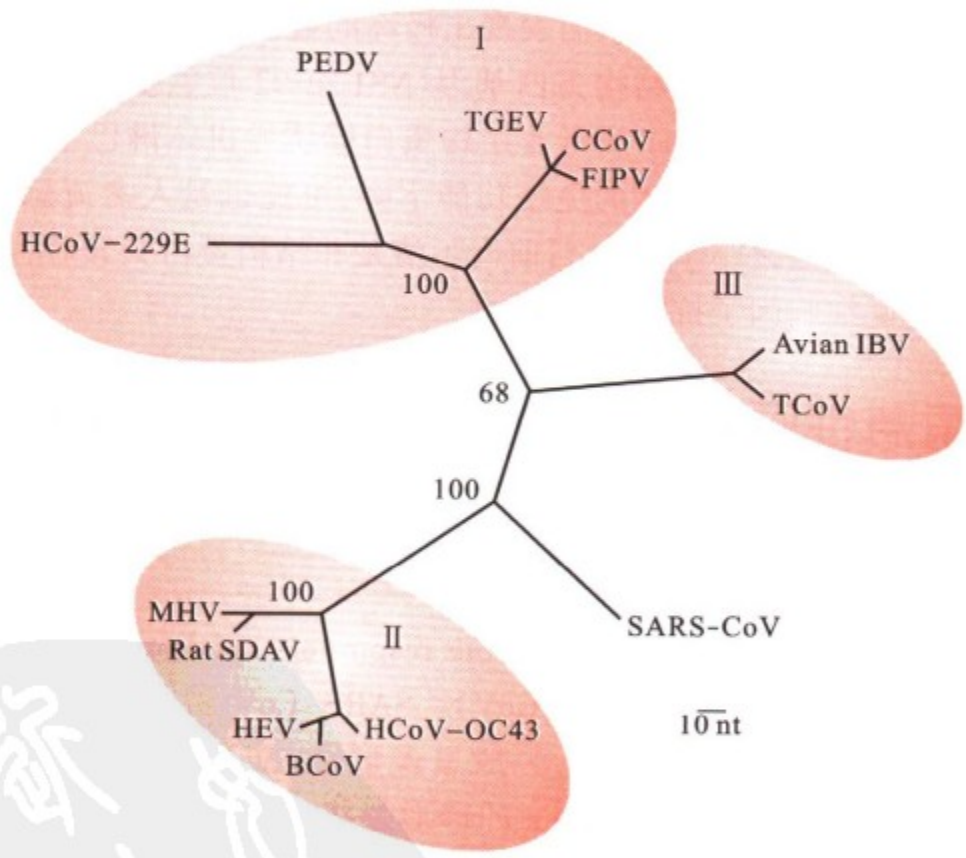



图 9-29 SARS-CoV 病毒与其他冠状病毒的系统进化分析

但也有人认为 SARS-CoV 可能是第 4 组冠状病毒的代表。一般认为 SARS-CoV 最可能来源于动物,特别是野生动物如果子狸、蝙蝠、猴、蛇等。

## 9.5.2 SARS-CoV 的基因组结构

冠状病毒是最大的 RNA 病毒,其基因组为单股正链 RNA。SARS-CoV 基因

组全长 27~30 kb,其 5′端有甲基化帽状结构,前约 2/3 的区域编码病毒 RNA 聚合酶复合蛋白,后 1/3 的区域编码病毒的结构蛋白和非结构蛋白,按基因组上的排列顺序依次为 S、E、M、N 蛋白，在 S 和 E 之间、M 和 N 之间以及 N 蛋白基因的下游有一些未知功能的可译框架,其 3′端有不少于 50 个碱基的 polyA 尾巴，这种结构与真核生物 mRNA 非常相似，这也是其基因组 RNA 自身即可在人体细胞内发挥翻译模板作用的重要结构基础。冠状病毒整个结构蛋白 mRNA 不存在转录后修饰、剪切等过程,而是直接在初级转录过程中,通过 RNA 聚合酶和某些转录因子以一种不连续转录机制，通过识别特定的转录调控序列(transcription regulating sequences,TRS) 有选择地从反义链 RNA 上一次性转录得到整个功能 mRNA。SARS-CoV 全基因组共有 14 个可译框架。在 SARS-CoV 的包膜上有 3 种主要糖蛋白,即 S 蛋白、M 蛋白和 HE 蛋白。S 蛋白被认为可能是病毒诱发机体免疫系统产生特异性抗体和细胞毒性 T 细胞的主要抗原蛋白,参与 SARS-CoV 侵入宿主细胞的过程。S 蛋白可进一步分为 S1 和 S2 两部分,S1 蛋白主要与宿主细胞膜受体结合,S2 蛋白介导病毒囊膜和宿主细胞膜融合。

### 9.5.3 SARS-CoV 的侵染过程

SARS-CoV 的生活周期包括病毒侵染、复制和组装及分泌几个阶段。在 SARS-CoV 侵染靶细胞时,病毒首先通过病毒囊膜上的 S 蛋白与敏感细胞的受体结合,然后通过其编码蛋白的自身折叠与相互结合牵引将病毒囊膜和细胞膜拉近而发生融合,使得病毒的遗传物质以内吞作用的方式进入靶细胞。最近的研究发现,ACE2 (血管紧张素转换酶Ⅱ) 是 SARS 冠状病毒的一个主要受体。ACE2 为依赖于锌离子的金属肽酶，是已知的唯一一个在血压调节中起关键作用的血管紧张素转换酶(ACE)的同系酶,该酶可将血管紧张素Ⅱ转换成血管紧张素。ACE2 是 SARS 病毒进入细胞的主要靶点,因此 ACE2 成为对 SARS 进行药理干预的首选目标。SARS 冠状病毒 S 蛋白的 S1 亚单位既具有与靶细胞受体结合的功能,又具有中和抗原表位。所以,针对 S1 亚单位的特异性抗体能够中和 SARS 冠状病毒感染。

SARS-CoV 病毒侵入细胞后基因组被释放到细胞质中,其转录和复制在细胞质中完成。病毒首先直接从其基因组 RNA 翻译出 RNA 依赖的 RNA 聚合酶。该聚合酶随后会以病毒基因组 RNA 为模板合成负链 RNA,之后再以新合成的负链 RNA 为模板转录生成新的病毒基因组 RNA。同时该聚合酶还会以负链基因组 RNA 为模板从基因组 RNA 的多个位点起始转录，产生 5~7 个亚基因组 mRNA 组分。亚基因组 RNA 都具有 5′端甲基化帽状结构和 3′端 poly A 尾结构,且长度呈递减分布。除最短的 mRNA 为单顺反子外其余的均为多顺反子,可以合成多聚蛋白前体,这些多聚蛋白再由病毒和宿主的蛋白切割,最终形成大小不同的功能蛋白质。

SARS-CoV 病毒的囊膜蛋白在粗面内质网内翻译,同时发生 N-糖基化,并在粗面内质网的表面聚合形成非共价的同源三聚体。随后被运输到高尔基复合体中进行高甘露糖的低聚糖基化修饰,同时部分区域连接末端糖基并发生脂酰化。成熟的囊膜蛋白主要集中在高尔基体,少量可以被转运至细胞膜表面引起

被感染细胞间发生膜融合而形成合胞体。在病毒蛋白合成的同时,病毒基因组进行复制,当病毒蛋白和基因组完成翻译和复制时,子代基因组 RNA 通过基因组中的包装信号与新合成的 N 蛋白结合,形成螺旋状的核衣壳之后再通过 M 蛋白的作用与囊膜结组装出芽。在病毒粒子的出芽过程中,M、S、E 蛋白嵌入到病毒粒子的囊膜中,形成的子代病毒粒子经细胞膜融合而释放到细胞外完成其生命周期(图 9-30)。

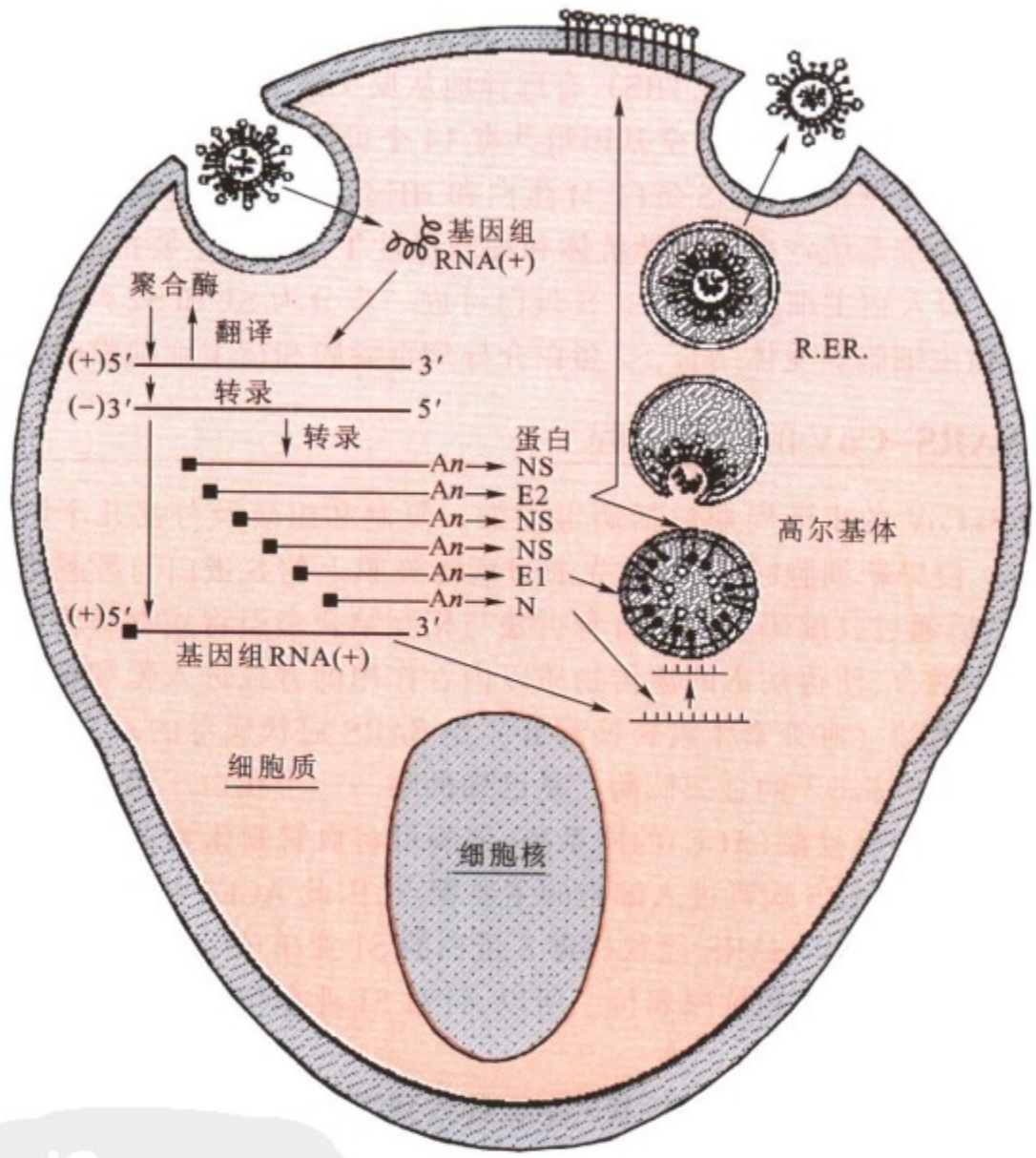



图 9-30 冠状病毒的生活周期图示

## 9.5.4 SARS-CoV 的起源及变异

目前,SARS-CoV 的起源仍不明了。但是,我国广东的一项调查显示,70%的果子狸携带 SARS-CoV,而 40%的野生动物经销商体内都含有 SARS-CoV 的抗体,其中单营果子狸者 SARS 冠状病毒感染率为 58%,明显高于单营蛇类者,暗示果子狸可能是 SARS 病毒的重要载体。基因组比较结果表明,果子狸的 SARS 样病毒比人类的 SARS-CoV 多 29 个核苷酸。从进化的角度看,果子狸 SARS 样病毒比人的 SARS-CoV 更原始,所以认为动物的 SARS 样病毒是人类 SARS-

CoV 的前体。有研究表明,从蝙蝠、猴、果子狸、蛇、老鼠和禽体内都可扩增出与人类 SARS 病毒高度同源的基因片段。因此,人 SARS-CoV 的最终源头问题有待于进一步研究。

## 9.6 基因治疗

人类疾病的发生,其实都是人体细胞中自身基因的改变或由外源病原体的基因产物与人体基因相互作用的结果。因此,长期以来,科学家设想人类能否最终运用遗传物质,无论是人类自身的或是外源遗传物质来治疗疾病,纠正人体本身基因结构或功能上的错乱,阻止病菌的侵染,杀灭病变的细胞或抑制外源病原体遗传物质的复制,保证人体健康。

### 9.6.1 基因治疗的主要途径

1990 年,科学家第一次用反转录病毒为载体,把腺苷脱氨酶基因(ADA)导入来自患者自身的 T-淋巴细胞,经扩增后输回患者体内,获得了成功。5 年后,患者体内 10%的造血细胞呈 *ADA* 基因阳性,除了还需服用小剂量的 ADA 蛋白之外,其他体征都正常。这一成功标志着基因治疗时代的开始。那么,基因治疗为什么具有诱人的前景,它与基因工程究竟有什么差别呢?

基因工程是将具有应用价值的基因,即“目的基因”,装配在具有表达元件的特定载体中,导入相应的宿主如细菌、酵母或哺乳动物细胞中,在体外进行扩增,经分离、纯化后获得其表达的蛋白产物。基因治疗是将具有治疗价值的基因,即“治疗基因”装配于带有在人体细胞中表达所必备元件的载体中,导入人体细胞,直接进行表达。进行基因治疗时无需对表达产物进行分离纯化,因为人细胞本身可以完成这个过程。

基因工程的“目的基因”主要是可分泌蛋白,如生长因子、多肽类激素、细胞因子、可溶性受体等,而非分泌性蛋白,如受体、各种酶、转录因子、细胞周期调控蛋白、原癌基因及抑癌因子等,由于它们不能有效地进入细胞而不能被应用于基因工程。基因治疗却不受上述限制。几乎所有的基因,只要它具有治疗作用,理论上均可应用于基因治疗。

基因工程的操作全部在体外完成,基因治疗则必须将基因直接导入人体细胞。这不仅在技术上具有很大难度,而且在有效性与安全性方面提出了更为苛刻的要求。

基因治疗主要有 *ex vivo* 和 *in vivo* 两条途径。

1. *ex vivo* 途径

这是指将含外源基因的载体在体外导入人体自身或异体(异种)细胞,这种细胞被称为“基因工程化的细胞”,经体外细胞扩增后输回人体。这种方法易于操作,而且因为细胞扩增过程中对外源添加物质的大量稀释,不容易产生不良反应。同时,治疗中用的是人体细胞,尤其是自体细胞,安全性好。但是,这种方法不易形成规模,而且必须有固定的临床基地。

2. *in vivo* 途径

这是将外源基因装配于特定的真核细胞表达载体上,直接导入人体内。这种载体可以是病毒型或非病毒型,甚至是裸 DNA。这种方式非常有利于大规模工业化生产。但是,对这种方式导入的治疗基因以及其载体必须证明其安全性,而且导入体内之后必须能进入靶细胞,有效地表达并达到治疗目的。因此,在技术上要求很高,其难度明显高于 *ex vivo* 模式。

## 9.6.2 基因治疗中的病毒载体

用于基因治疗的病毒载体应具备以下基本条件:

① 携带外源基因并能装配成病毒颗粒。

② 介导外源基因的转移和表达。

③ 对机体没有致病力。

因为大多数野生型病毒对机体都具有致病性,需要对其进行改造后才能被用于人体。理论上,各种类型的病毒都能被改造成病毒载体,但由于病毒的多样性及与机体之间复杂的依存关系,人们至今对许多病毒的生活周期、分子生物学、与疾病发生发展的关系等的认识还很不全面,从而限制了许多病毒发展成为具有实用性的载体。到目前为止,只有少数几种病毒如反转录病毒、腺病毒及腺病毒伴随病毒、疱疹病毒(包括单纯疱疹病毒、痘苗病毒及 EB 病毒)等被成功地改造成为基因转移载体。

1. 病毒载体的产生

研究病毒载体首先要对病毒的基因组结构和功能有充分的了解,最好在获得病毒基因组全序列信息的基础上进行。病毒基因组可分为编码区和非编码区,编码区的序列产生病毒的结构蛋白和非结构蛋白。根据其对病毒感染性复制的影响,又可分为必需基因和非必需基因。非编码区中含有病毒进行复制和包装等功能所必需的顺式作用元件。各种野生型病毒颗粒都具有一定的包装容量,即对所包装的病毒基因组的长度有一定的限制。一般来说,病毒包装容量不超过自身基因组大小的 105%~110%。

最简单的办法是将适当长度的外源 DNA 插入病毒基因组的非必需区,包装成重组病毒颗粒。有科学家将 4.5kb 的 *lacZ* 基因表达盒插入 HSV-1 病毒的 *UL44*(编码糖蛋白 C)基因的 *Xba* Ⅰ位点中,构建成重组 HSV 病毒。由于 *UL44* 基因产物对于 HSV 在培养细胞中产毒性感染是非必需的,因此,该重组病毒可以在细胞中增殖传代。用这种重组病毒感染细胞,发现 *lacZ* 基因能进入细胞并得到高效表达。

然而,这样的重组病毒作为基因转移载体有许多缺点。首先,许多野生型病毒可通过在细胞中的复制而导致细胞裂解死亡,或带有病毒癌基因而使细胞发生转化,必须经过改造使其成为复制缺陷型病毒并且删除致癌基因后才能用于基因治疗。其次,插入外源 DNA 的长度受到很大限制,尤其对于基因组本身较小的病毒如腺病毒伴随病毒(4.7 kb)、反转录病毒(8~10 kb)、腺病毒(36 kb),如果不去除病毒基因,可供外源 DNA 插入的容量就很小,要删除更多的病毒基

因以腾出空间以插入较大的外源 DNA。为了增加病毒载体插入外源 DNA 的容量,除了删除病毒的非必需基因外,还可以进一步删去部分或全部必需基因,这些必需基因的功能由辅助病毒或宿主细胞提供。

2. 病毒载体的分类

(1) 重组型病毒载体　这类载体以完整的病毒基因组为改造对象,在不改变病毒复制和包装所需的顺式作用元件的情况下,有选择性地删除病毒的某些必需基因,尤其是前早期或早期基因以控制其表达,所缺失的必需基因的功能由同时导入细胞中的外源基因表达单位提供。一般通过同源重组方法将目的基因插入到病毒基因组中。

(2) 无病毒基因的病毒载体　这类载体可以看做是重组病毒载体的一种极端减毒情况,往往由重组载体和辅助系统组成。重组载体主要由外源基因表达盒、病毒复制和包装所必需的顺式作用元件及载体骨架组成。辅助系统包括病毒复制和包装所必需的所有反式作用元件。在辅助系统的作用下,重组载体以特定形式(单链或双链 DNA 或 RNA)被包装到不含有任何病毒基因的病毒颗粒中。这类载体的优点在于载体病毒本身安全性好,容量大。缺点在于往往需要辅助病毒参与载体 DNA 的包装,造成终产品中辅助病毒污染,影响其应用。

3. 病毒载体在基因治疗中的应用

据不完全统计,自 1990 年开展第一例人体基因治疗以来,已进行了数百项基因治疗临床试验,涉及患者上千人,其中 2/3 以上的临床方案应用了病毒载体进行基因导入。不同的病毒载体具有特定的生物学功能,决定了其在基因治疗中的应用,表 9-9 列举了常用的病毒载体的特性和适用范围。

表 9-9　常用病毒载体的特性和适用范围

| 病毒载体 | 生物学特性 | 适用范围 |
|---|---|---|
| 反转录病毒载体 | 可感染分裂细胞 | *ex vivo* 基因治疗 |
| | 整合到染色体中 | |
| | 表达时间较长 | 肿瘤基因治疗 |
| | 有致癌的危险 | |
| 腺病毒载体 | 可感染分裂和非分裂细胞 | *in vivo* 基因治疗 |
| | 不整合到染色体中 | |
| | 外源基因表达水平高 | 肿瘤基因治疗 |
| | 表达时间较短 | 疫苗 |
| | 免疫原性强 | |
| 腺病毒伴随病毒载体 | 可感染分裂和非分裂细胞 | *in vivo* 基因治疗 |
| | 整合到染色体中 | |
| | 无致病性,免疫原性弱 | *ex vivo* 基因治疗 |
| | 可长期表达外源基因 | |
| | 在骨骼肌、心肌、肝、视网膜等组织 | 遗传病基因治疗 |

续表

| 病毒载体 | 生物学特性 | 适用范围 |
| --- | --- | --- |
| | 中表达水平较高 | 获得性慢性疾病的基因治疗 |
| 疱疹病毒载体 | 具嗜神经性,可逆轴突传递 | 神经系统疾病的基因治疗 |
| | 可潜伏感染 | |
| | 容量大 | 肿瘤的基因治疗 |
| | 可感染分裂和非分裂细胞 | |

4. 基因治疗中的问题

基因治疗要取得突破,必须在基因导入系统、基因表达的可控性及获得更多更好的治疗基因这三个方面下工夫。

(1) 靶向性基因导入系统　基因治疗中的关键问题是必须将治疗基因送入特定的靶细胞,并在该细胞中得到高效表达。这对于恶性肿瘤治疗尤为重要,如果不能有效地将治疗基因导入大多数肿瘤细胞,则至少要求它尽可能不进入或较少进入正常细胞。因此,目前针对恶性肿瘤的免疫基因治疗仍按 *ex vivo* 形式操作,很少有体内直接导入的治疗模式。病毒型载体中,除直接注射入瘤体外,若用全身给药,在肿瘤细胞中分布极低,很难期望达到治疗作用。当务之急,是要尽快建立靶向性导入系统。

(2) 外源基因表达的可控性　由于电脉冲介导裸 DNA 技术的改进,科学家已能将许多基因,如分泌性蛋白基因(包括生长因子、激素、细胞因子,可溶性受体)导入肌肉并维持相当时间的表达,已达到可发挥药效的水平。但是,如果这些基因导入后表达处于无调控状态,将会造成严重后果。

最理想的可控性是模拟人体内基因本身的调控形式。这是今后长期的追求目标,需要全基因或包括上下游的调控区及内含子序列,将对导入基因的载体系统产生严峻的挑战,因为今后设计的载体需有几十 kb 甚至上百 kb 的包装能力。表 8-9 中列举了已报道的诱导系统,其基本原理如图 9-31,其中大部分仅在体外有效。以酵母 *gal* 4 系统为例,在目的基因上游区接上 *gal* 4 的顺式元件,所用的激活蛋白是一个带有疱疹病毒 VP16/*gal* 4 DNA 结合区的孕酮受体的变异体。该变异体(PR-LBD△)只能与孕酮的拮抗剂(RU486)结合而不与孕酮或其他衍生物结合。当体系中不存在 RU486 时,治疗基因的表达水平极低;而给予 RU486 后,RU486 与 PR-LBD△结合,激活 VP16,杂种激活蛋白形成二体并通过 *gal* 4 DNA 结合区与治疗基因上游区的 *gal* 4 顺式元件结合,启动治疗基因的表达。

(3) 治疗基因过少　目前用于临床试验的治疗基因数量很少。绝大部分多基因疾病,如恶性肿瘤、高血压、糖尿病、冠心病、神经退行性疾病的致病基因还有待阐明,因此,可选择的靶基因不多。只有尽快发现并克隆大量的功能基因,迅速确定其调控序列,基因治疗才能获得更大的发展。

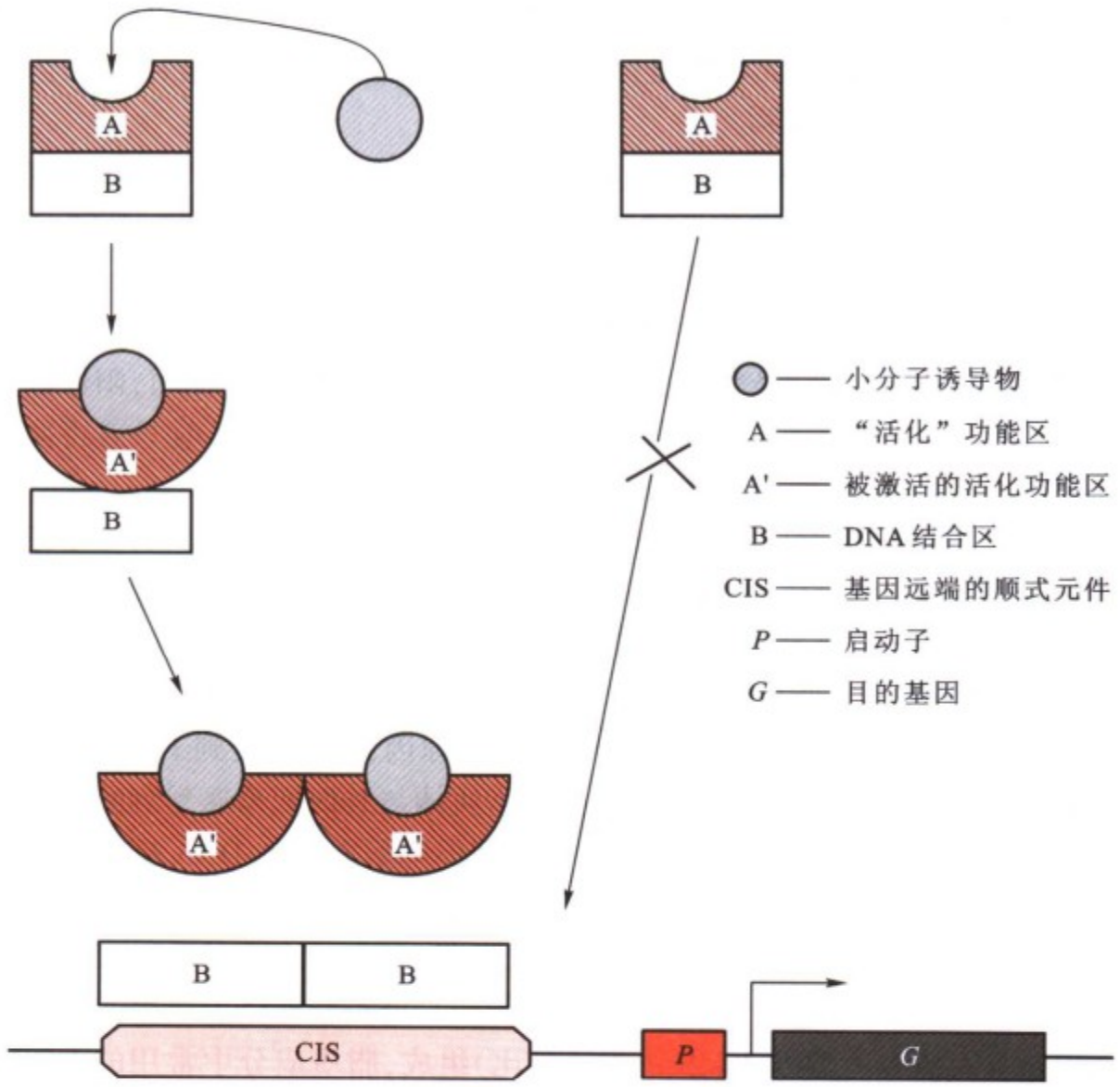



图 9-31 导入基因的表达诱导框架图

## 9.6.3 基因治疗中的非病毒载体

病毒载体存在许多不足，主要体现在免疫原性高、毒性大、目的基因容量小、靶向特异性差、制备较复杂及费用较高等。现在，人们愈来愈重视人工合成的非病毒载体的研究。目前常用的非病毒载体包括裸 DNA(naked DNA)、脂质体载体(liposome vector)及阳离子多聚物型载体(cationic polymer vector)等。

1. 裸 DNA

裸 DNA，又称自由 DNA，是结构最简单的非病毒载体。使用裸 DNA 进行基因治疗的最大困难在于如何将目的基因导入相应的细胞并得到有效而长期的表达。主要通过物理或机械方法(如直接注射或基因枪法)将 DNA 导入适当部位(如皮肤、骨骼肌、肝、支气管内、心肌和瘤体内)。当用于激发免疫反应时又称为 DNA 疫苗(DNA vaccine)，是一种非常有希望的疫苗方式。

基因枪方法又叫微粒轰击技术，它采用能自发吸收 DNA 的钨或金微粒，通过高压电所产生的高能电弧促使被 DNA 包裹着的金属颗粒产生极高的速度，有效穿透单细胞层或靶器官从而将 DNA 导入培养细胞或动物组织中。

裸 DNA 虽能有效运转并表达目的基因，但缺乏靶向性，并且只能在局部作用，不能转染大量细胞，经常需要进行外科手术以暴露靶器官，使用时局限性较大。

2. 脂质体/DNA 复合物

脂质体是具有双层膜的封闭式粒子，它们能促进极性大分子穿透细胞膜。根据脂质体包裹 DNA 的方式不同可将脂质体分为阳离子脂质体、阴离子脂质体、pH 敏感脂质体及融合脂质体等。

阳离子脂质体是目前最常用的脂质类型，它主要由带正电荷的脂质及中性辅助脂质组成，两者通常是等物质的量混合。由于电荷的相互作用，阳性电荷的脂质体和带负电荷的 DNA 之间可以有效地形成复合物，制备时只需将两者直接混合形成复合物，不受基因体积大小的限制。

脂质的化学结构、组成对载体的活性和毒性有很大的影响。具有多个正电荷头部的脂质体转染率较高，这可能是因为它与 DNA 结合更牢固。脂质体在酸性环境中与细胞膜结合能力大大降低，在细胞间的转运也取决于脂质体中类脂的成分。DC-Chol/DOPE 阳离子脂质体是第一个被批准用于人体的脂质体。由于制备技术的不稳定性，对靶细胞的毒性和试剂价格较贵等原因，其应用受到了一定的限制。

细胞摄取脂质体主要通过内吞途径，脂质体与其内含物遇到溶酶体即被降解。pH 敏感脂质体可在一定程度上避免溶酶体降解并增加 DNA 摄取量和稳定性。pH 敏感脂质体的原理是酸性条件下脂肪酸羧基质子化，形成六角晶相发生膜融合。在形成内吞小泡进入溶酶体之前，由于 pH 从 7.4 降至 5.3~6.3，pH 敏感脂质体与内吞小泡膜融合，将内含物导入胞质。这类脂质体主要由二油酰胆碱(DOPE)、胆固醇和油酸以 4:4:2(物质的量比)组成，脂质成分中常用的 DOPE 决定了脂质体在中性条件下的稳定性和酸性条件下与细胞融合的能力。

pH 敏感脂质体虽然通过静电吸附与细胞膜产生非特异性相互作用，转运效率比普通脂质体高，但实际上也只有 10%的基因释放至细胞质中，其余仍在溶酶体中降解。

3. 多聚物/DNA 复合物

由于 DNA 带负电，利用阳离子多聚体的氨基基团的正电荷与 DNA 的磷酸基团结合发生电性中和，可以使 DNA 缩合形成稳定的多聚复合物(polyplex)，不易被核酸酶降解，并可防止沉淀，从而提高转染效率。此外，复合体大小约 80~100 nm，带正电荷，可与细胞表面带负电荷的受体结合，因此能有效地被内吞摄入，介导基因转移。常见的阳离子多聚体有多聚左旋赖氨酸、鱼精蛋白、组蛋白、多聚乙胺、多聚乙烯亚胺和星状树突体等，可以分别形成线型、分枝型、球型或类球体状结构。虽然多聚阳离子型载体系统具有合成比较方便、安全无毒的特点，但是，单独使用多聚阳离子同样具有基因转移组织特异性和靶向性差、体内基因转移效率较低等缺点，有待于进一步改善和提高。

4. DNA 微球体(DNA-nanospheres)

白明胶和聚氨基葡萄糖等生物高聚物属阳离子复合物，能与 DNA 形成大小为 200~750 nm 的微球体复合物。白明胶和聚氨基葡萄糖等有如下特点：①它们是可降解生物材料；②DNA 可均一的分布在微球体中，免于被 DNA 酶降解；

③可与活性物质共包装达到靶向运输增加了基因治疗的可能性;④储存方便稳定;⑤可以达到可控持续释放,延长了基因表达时间;⑥毒性低。

## 思考题

1. 癌症已经成为威胁人类健康的主要杀手,根据你所了解的知识,简述癌症为何具有如此大的危害? 对于癌症的防治为何如此困难?
2. 如果你是一家生物技术公司的CEO或者负责技术的高层人员,有人向你推荐一种治疗癌症的特效药,寻求与你们公司合作。你将在哪些方面对这种药物进行评价?
3. HIV是通过攻击人体内T4-淋巴细胞,使人体的免疫系统遭到破坏而致病的。能不能对艾滋病患者进行定时输入通过基因工程生产的特定T4-淋巴细胞而延长其生命? 主要存在哪些技术上的障碍?
4. 人类为防治HIV作了大量的工作,但是HIV用高的变异速度以逃避人体的免疫系统和药物,为什么HIV的变异如此之快? HIV的变异有无规律? 主要在哪些方面发生变异?
5. 不同人感染HIV和HBV在人体内潜伏的时间存在很大差异,有哪些原因?
6. HIV最初的宿主是猩猩,而猩猩与人类的基因组几乎完全相同。问为何HIV在这两种生物中有不同的危害? 人类是否可借助猩猩抵御HIV的方式防治HIV?
7. 比较SARS-CoV引起的非典和人禽流感,哪种病在流行上易于控制? 为什么?
8. 病毒致癌基因很多是宿主本身的原癌基因,这种选择对于病毒本身有何意义?
9. 基因治疗比较难于推广应用的主要原因是载体。简述病毒载体和非病毒载体的优缺点并加以比较。
10. 假如你得到一个据说通过改造并适合作为基因治疗载体的病毒,你将如何分析这种可行性? 主要在哪些方面作改造? 你将如何测试改造后的病毒载体的效率和安全性?

## 主要参考文献


1. Ksiazek T G, Erdman D, Goldsmith C S, *et al.* A novel coronavirus associated with severe acute respiratory syndrome. New Engl. J. Med. 2003, 348: 1953–1966.
2. Ye Q Z, Krug, R M, Tao Y J. The mechanism by which influenza A virus nucleoprotein forms oligomers and binds RNA. Nature, 2006, 444: 1078–1082.
3. Noda T, Sagara H, Yen A, *et al.* Architecture of ribonucleoprotein complexes in influenza A virus particles. Nature, 2006, 439: 490–492.
4. Telenti A, Goldstein D B. Genomics meets HIV-1. Nature Reviews Microb. 2006, 4: 865–873.
5. Nabel G J. Close to the edge: neutralizing the HIV-1 envelope.Science, 2005, 308: 1878–1879.
6. Zhou T, Xu L, Dey B, *et al.* Structural definition of a conserved neutralization epitope on HIV-1 gp120.Nature, 2007, 445: 732–737.

# 10 基因与发育


10.1 免疫系统发育及免疫球蛋白基因表达
10.2 果蝇的发育与调控
10.3 高等植物花发育的基因调控


## 10.1 免疫系统发育及免疫球蛋白基因表达

免疫系统是动物的主要防御系统,包括淋巴细胞(lymphocyte,分 B-淋巴细胞和 T-淋巴细胞)及组织相容性抗原两部分。另外,巨噬细胞(macrophage)也参与免疫反应。B-淋巴细胞由骨髓产生,主要功能是制造用于对付外来抗原的抗体,它专一地识别相应的抗原。T-淋巴细胞产生于胸腺(thymus),它的主要功能是产生 T-淋巴细胞受体,识别产生抗体的 B-淋巴细胞,以帮助巨噬细胞发挥功能。生物有机体还具有一套特定的组织相容性抗原(histocompatibility antigen),保证机体根据组织相容性抗原的类型识别自身细胞和异源细胞,参与淋巴细胞及其他许多细胞间的反应。

抗体、T-淋巴细胞受体和组织相容性抗原是免疫系统中 3 种最重要的成分。这 3 类蛋白质有一个共同的特点:多样性。由于每类蛋白质都由一个巨大的基因家族编码,它们各自具有许多变异型。在抗体和 T-淋巴细胞受体中,基因片段重排导致了所编码蛋白质的多样性。

### 10.1.1 脊椎动物免疫系统

脊椎动物免疫系统产生的抗体能识别和结合数百万种潜在的"非本身"的抗原分子(antigen),这意味着免疫系统细胞要合成数量巨大的抗体分子(antibody)。因为抗体是一种蛋白质,它的结构由相应的基因决定,数目巨大的抗体分子是否来自相应数目的基因呢?事实上,生物体内参与抗体合成的基因数目远小于抗体数

目。那么,少量的基因通过什么样的调控机制表达了数目巨大的抗体分子呢?脊椎动物抗体基因的表达与多样性问题是近 20 年来分子生物学界研究的中心议题之一。

生物体内存在着两条不同的免疫途径:一是体液免疫(humoral immunity),二是细胞免疫(cellular immunity)。在循环体系免疫中,B-淋巴细胞首先识别外源致病因子或其他任何形式的抗原,通过浆细胞(即成熟 B-淋巴细胞)分泌出能溶于血液蛋白质中的免疫球蛋白(immunoglobulin,Ig)。这些被称作抗体的 Ig 分子通过特定的结合位点形成抗原-抗体复合物,最终被巨噬细胞所吞噬。Ig 可占血液总蛋白的 20%。在细胞水平免疫体系中,抗原受体细胞产生出类似 Ig,紧密结合在细胞膜表面的蛋白质分子。这些抗体分子能与 T-淋巴细胞表面的受体专一结合,最终导致整个感染细胞被后者降解(图 10-1)。

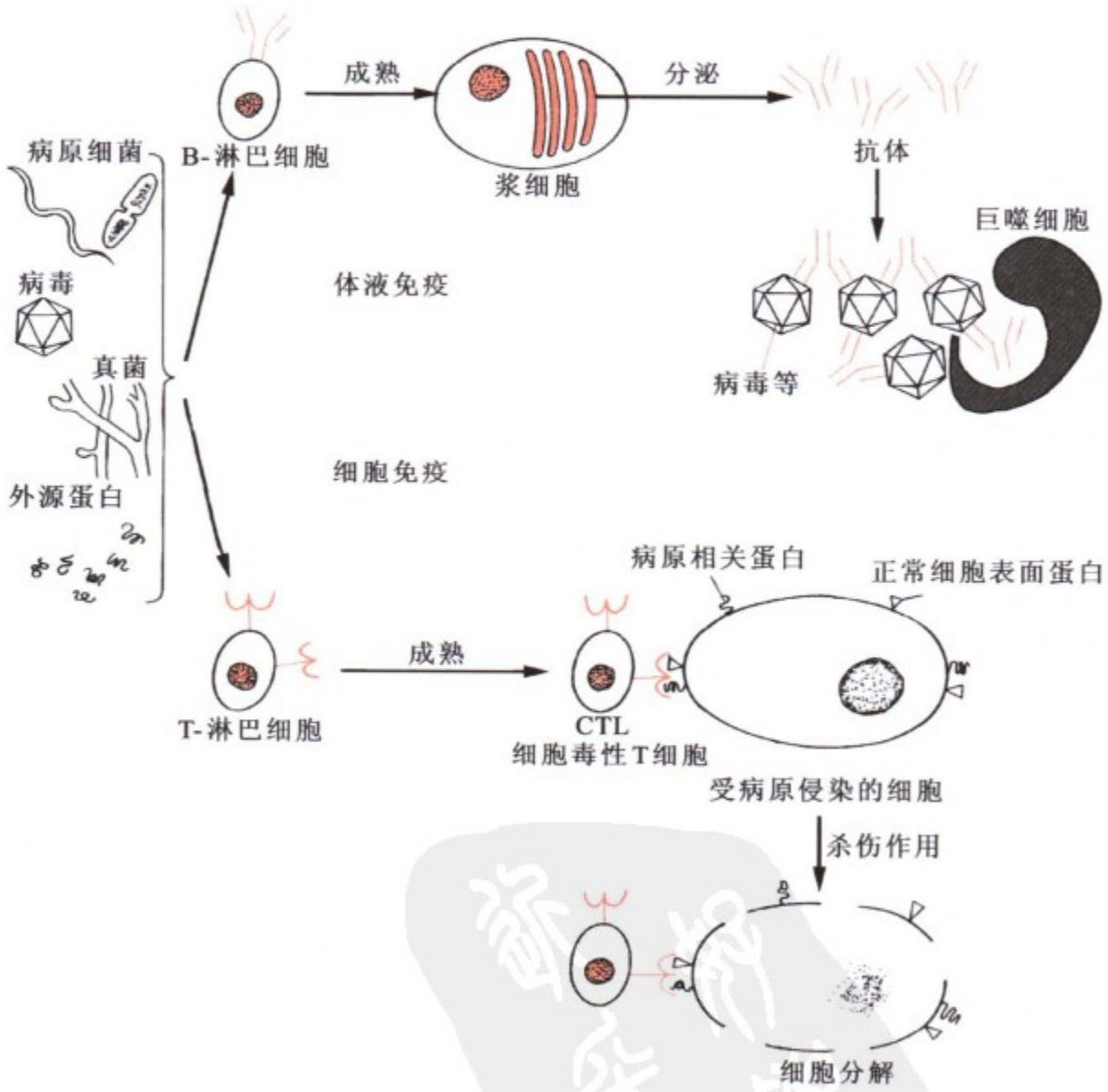



图 10-1 生物体内存在的两条主要免疫途径:体液免疫和细胞免疫图示

## 10.1.2 B-淋巴细胞的发育和分化

B-淋巴细胞(B lymphocyte),简称 B 细胞。人和其他哺乳动物 B 细胞分化产生于骨髓中(图 10-2)。成熟的 B 细胞主要存在于淋巴结皮质浅层的淋巴小

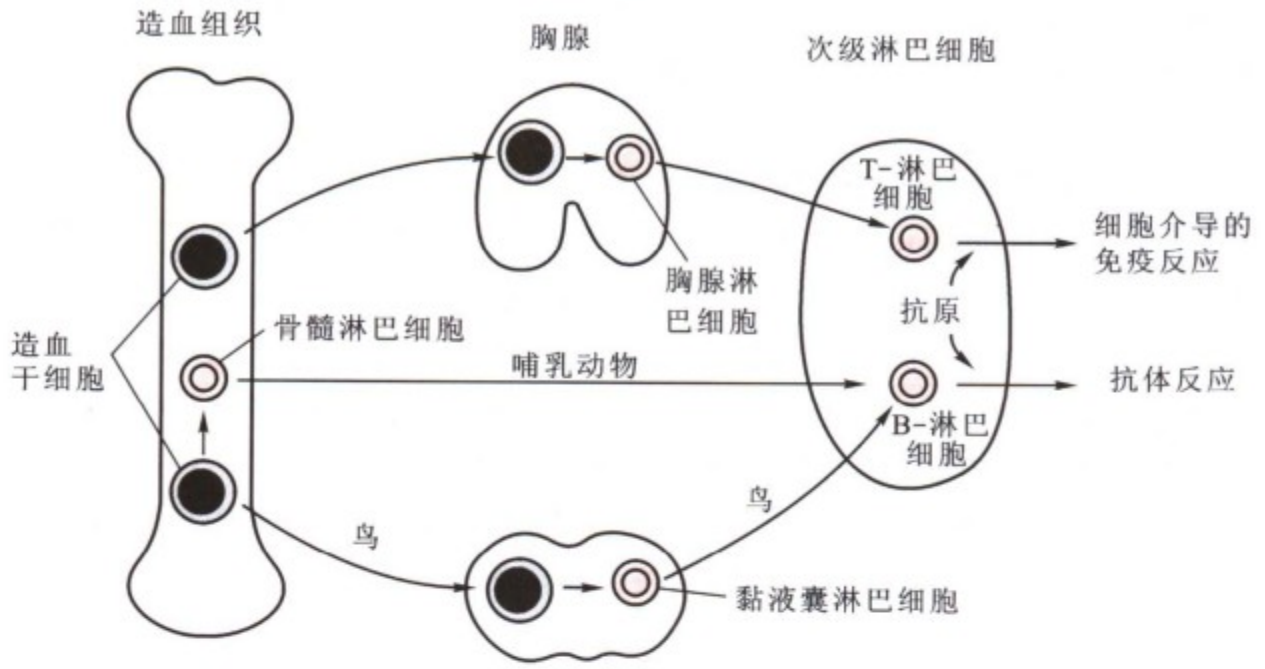



图 10-2 B-淋巴细胞和 T-淋巴细胞的发育过程

结及脾脏的红髓与白髓淋巴小结内。B 细胞在抗原刺激下可分化为浆细胞，合成和分泌免疫球蛋白来专一地识别相应的外来抗原，主要执行机体的循环免疫功能。

1. B 细胞的分化

哺乳动物 B 细胞的分化过程可分为以下几个阶段：祖 B 细胞、前 B 细胞、未成熟 B 细胞和成熟 B 细胞等。其中祖 B 细胞、前 B 细胞、未成熟 B 细胞的分化不依赖于抗原，其分化过程在骨髓中完成（图 10-3）。抗原依赖阶段是指成熟 B 细胞受抗原刺激活化并分化为浆细胞的过程，主要在外周免疫器官中完成。祖 B 细胞（pro-B cell）由骨髓淋巴细胞分化而来，仅存在于骨髓、胎肝等组织中。在早期祖 B 细胞中，*D-J* 基因片段开始重排，并开始表达 B 细胞受体（B cell receptor，BCR）复合物中的 Igα 和 Igβ。在晚期祖 B 细胞中 *V-DJ* 发生重排。根据

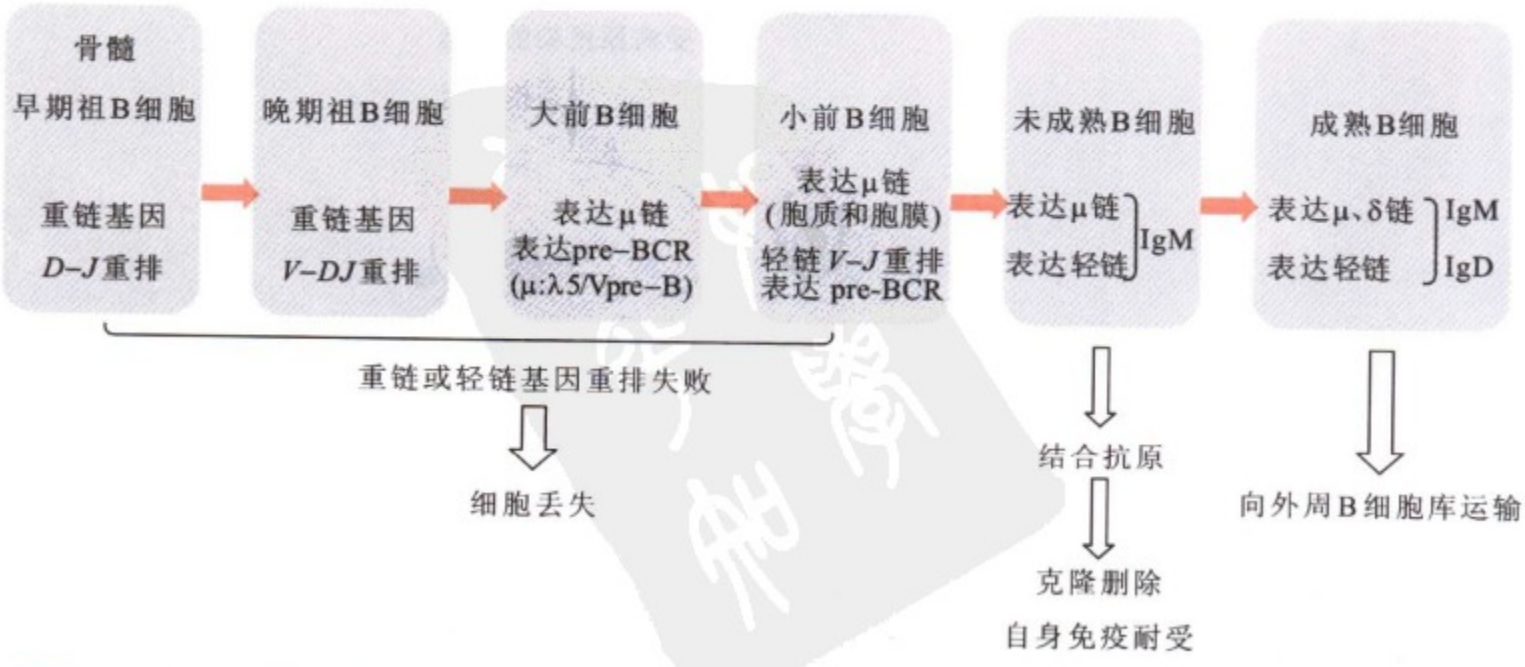



图 10-3 B-淋巴细胞的发育和分化

前 B 细胞的形态以及轻链基因的重排，可将其分为大前 B 细胞和小前 B 细胞。大前 B 细胞 *V-DJ* 重排后表达 μ 链，并表达前 B 细胞受体。小前 B 细胞轻链只发生 *V-J* 发生重排，无功能性 Ig 表达，缺乏对抗原的反应能力。未成熟 B 细胞开始表达 IgM，此时如与抗原相结合，则产生负应答，发生克隆删除，是形成自身免疫耐受的机制之一。成熟 B 细胞受相应抗原刺激后发生增殖和分化，并在抗原呈递细胞和 Th 细胞的辅助下成为活化 B 细胞，进而分化为浆细胞(plasma cell)或称抗体分泌细胞，合成和分泌各种免疫球蛋白。此时，膜表面 Ig 水平逐渐降低，分泌型 Ig 逐渐增加，并可发生免疫球蛋白基因轻链的类型转换。活化 B 细胞中的一部分可转变为小淋巴细胞，又称为记忆 B 细胞 (memory B cell)，停止增殖和分化，并可存活数月至数年。一旦细胞再次与同一抗原相接触，便很快发生活化和分化，迅速产生较高水平的抗体，而且维持时间也较长。

2. B 细胞膜表面分子标记

B 细胞表面有多种膜表面分子，包括白细胞分化抗原、MHC 以及多种膜表面受体，这些分子可被用于识别抗原，与免疫细胞和免疫分子相互作用，同时也是分离和鉴别 B 细胞的重要依据。膜表面球蛋白是 B 细胞特异性识别抗原的受体，也是 B 细胞重要的特征性标志。未成熟 B 细胞表达 IgM，成熟 B 细胞同时表达 IgM 和 IgD，有的成熟 B 细胞还表达 IgG、IgA 和 IgE。B 细胞表面所有 Ig 的可变区都由相同的 $V_H$ 和 $V_L$ 基因所编码。抗原刺激后 B 细胞表面 IgD 很快消失，记忆 B 细胞表面也不存在 IgD(图 10-4)。

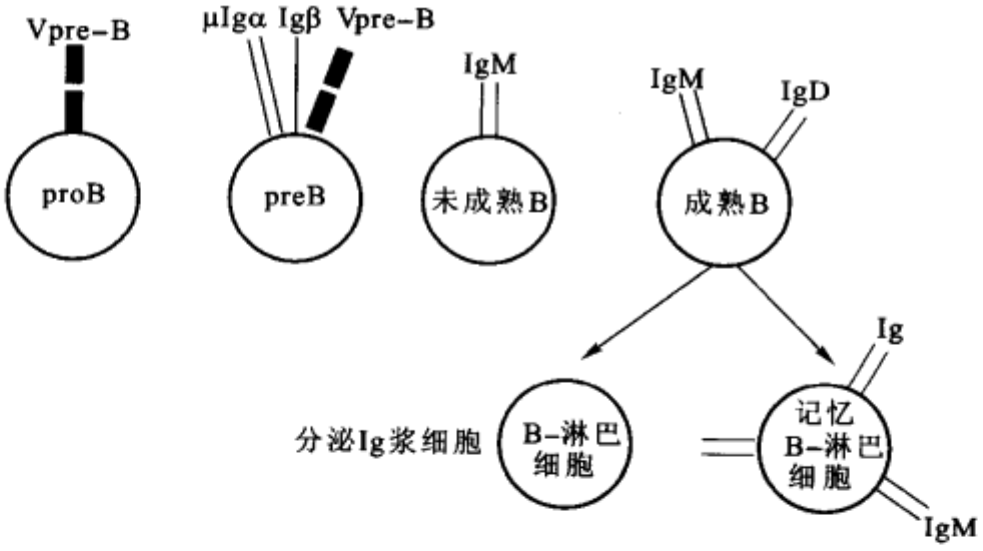



图 10-4 B 细胞发育和分化过程中有不同免疫球蛋白表达

B 细胞不仅表达 MHC Ⅰ类抗原，也能表达 MHC Ⅱ类抗原，表达时期集中在前 B 细胞至活化 B 细胞期。B 细胞表面的 MHC Ⅱ类抗原在 B 细胞与 T 细胞相互协作时起重要作用。B 细胞同时还具有抗原呈递作用。

3. B 细胞的功能

根据成熟 B 细胞表面是否表达 CD5 分子而将 B 细胞分为 B-1 ($CD5^+$)和 B-2($CD5^-$)两个亚群。B-1 细胞参与非特异性免疫或先天免疫，B-2 则是通常意义上的 B 细胞，主要介导特异性免疫中的循环体系免疫。B-1 细胞出现在个体发育过程中的早期，主要在胚胎的肝组织、视网膜中发育，少部分在胚胎出生前

的骨髓中发育。在成年人(或动物)骨髓中 B 细胞谱系中不存在 B-1 亚群,仅有 B-2 亚群。

B 细胞是免疫系统中产生抗体的主要细胞,也是专职的抗原呈递细胞,可有效摄取、加工和呈递可溶性抗原。此外,B 细胞还可分泌多种细胞因子,参与机体的免疫调节。B 细胞受到抗原刺激后活化、增殖、分化为浆细胞,然后合成各种类型的 Ig,并参与机体的循环体系免疫。非活化状态的 B 细胞不产生细胞因子,而活化的 B 细胞可产生多种细胞因子,参与免疫应答及机体的造血和炎症反应等。

### 10.1.3 T-淋巴细胞发育及分化

T-淋巴细胞(T lymphocyte)来源于骨髓的淋巴样干细胞,简称 T 细胞,主要功能是产生一种蛋白质,即 T-淋巴细胞受体。T 细胞不仅直接执行细胞免疫功能,而且还可以通过产生多种细胞因子或表达黏附分子等手段,与其他免疫细胞发生直接或间接的接触,从而发挥广泛的免疫调节作用。

1. T 细胞的分化与发育

研究小鼠胸腺细胞的发育过程表明,在胚胎发育到 11~12 d 时,淋巴干细胞进入胸腺,并在胸腺微环境的影响下迅速发生增殖和分化。诱导 T-淋巴细胞在胸腺内分化、成熟的主要因素有:①胸腺基质细胞(thymus stromal cell,TSC)能分泌多种细胞因子(IL-1/IL-6 和 IL-7)及胸腺激素,并通过其细胞表面的黏附分子直接与胸腺细胞发生相互作用,诱导后者分化;②胸腺中的“抚育细胞”(nurse cell)也参与调控 T 细胞的成熟和分化过程;③胸腺细胞自身可分泌 IL-2 和 IL-4 等多种细胞因子,影响本身的分化和成熟;④巨噬细胞、胸腺内上皮细胞和树突细胞也参与胸腺细胞分化过程中的自身耐受以及 T 细胞功能性亚群的形成。

2. Notch 信号途径与 T 细胞分化

在免疫系统发育过程中,Notch 信号途径使淋巴干细胞向 T 细胞前体分化,表达功能性 TCRαβ 而非 TCRγδ,并使 $CD4^+CD8^+$T 细胞向 $CD4^-CD8^+$T 细胞表型分化。胎肝或骨髓中的淋巴样干细胞进入胸腺,发育成为不成熟的 $CD4^-CD8^-$T 细胞和 $CD4^+CD8^+$T 细胞,再经过阳性选择,分化成 $CD4^-CD8^+$T 细胞或 $CD4^+CD8^-$T 细胞。这一发育和分化过程受到多种因素的严密调控,其中 Notch 信号发挥了关键性作用。该信号决定了骨髓淋巴样干细胞向 B 系或 T 系方向分化,调控了 $CD4^+CD8^+$ T 细胞的阳性选择过程,提高了胸腺中成熟的 $CD4^-CD8^+$ T 细胞比例,促进了 TCRαβ 型 T 细胞的形成。Notch 信号途径对淋巴干细胞分化和发育调控作用见图 10-5。

3. T 细胞表面受体

T 细胞受体(T cell receptor,TCR)是 T-淋巴细胞表面识别自身 MHC-抗原肽复合物的受体。TCR 在细胞表面与 CD3 组成 TCR-CD3 复合物,TCR 识别抗原后其刺激信号是由 CD3 分子传递的。在同种异体移植中,TCR 也识别单独的非己 MHC 抗原。

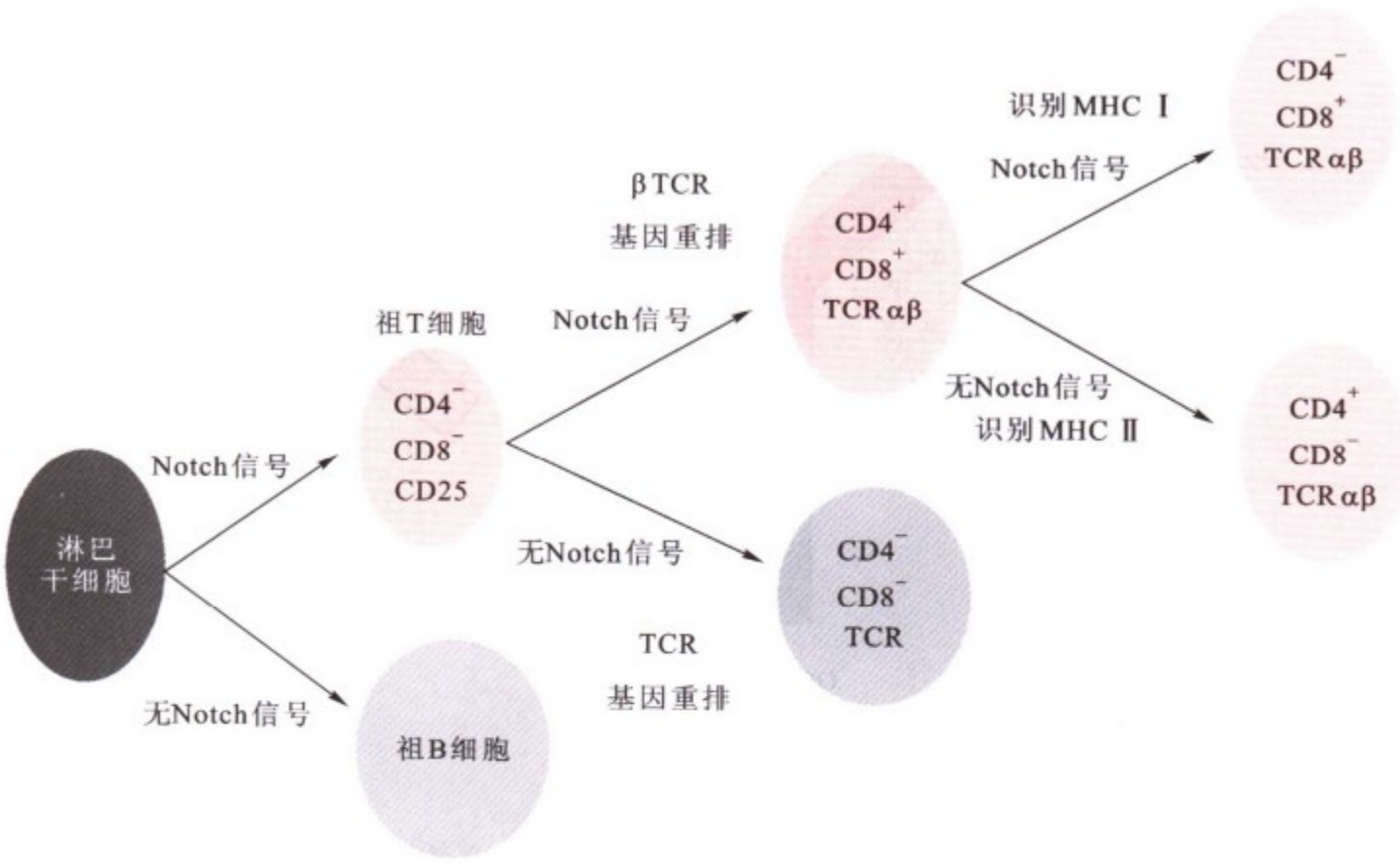



图 10-5 Notch 信号对淋巴细胞分化与发育的调控作用

TCR 中的多肽链组成是异质性的，根据抗原结构和编码基因不同，有 α、β、γ 和 δ 四种多肽，这四种多肽均属于免疫球蛋白基因超家族的成员。成熟的 TCR 多肽相对分子质量介于 40 000~60 000。根据 TCR 中异源双体组成上的差异，可以将 TCR 分为 TCRαβ 和 TCRγδ 两类。

### 10.1.4 免疫球蛋白的结构

抗体分子虽然是多种多样的，但其基本结构都是由一个四聚体的免疫球蛋白分子，由 4 条（两对）多肽链组成，包括两条相同的轻链（L-chain）和两条相同的重链（H-chain）。轻链和重链在相对分子质量上有较大差别（图 10-6），前者约 $2.3\times10^4$，后者则介于 $5.3\times10^4$~$7.0\times10^4$。每条链都有两个区，位于氨基端的是可变区（variable region，V 区）。不同的免疫球蛋白分子中，V 区的氨基酸序列是不同的。位于羧基端的是恒定区（constant region，C 区）。免疫球蛋白分子中，C 区的氨基酸序列十分保守。

所有 Ig 分子都含有两类轻链中的一类，即 κ 型或 λ 型。κ 型和 λ 型轻链的恒定区和可变区的氨基酸序列都是不同的。在小鼠中，95%的抗体轻链是 κ 型，而人类抗体轻链中，κ 型和 λ 型各占 50%左右。根据抗体的重链结构特征将其分为 5 类（表 10-1），再根据抗原的不同和结构的变化又可分出若干亚类。如人类有 4 种 IgG 亚类，2 种 IgA 亚类和 2 种 IgM 亚类，每一种类型都有一个特征性的恒定区。

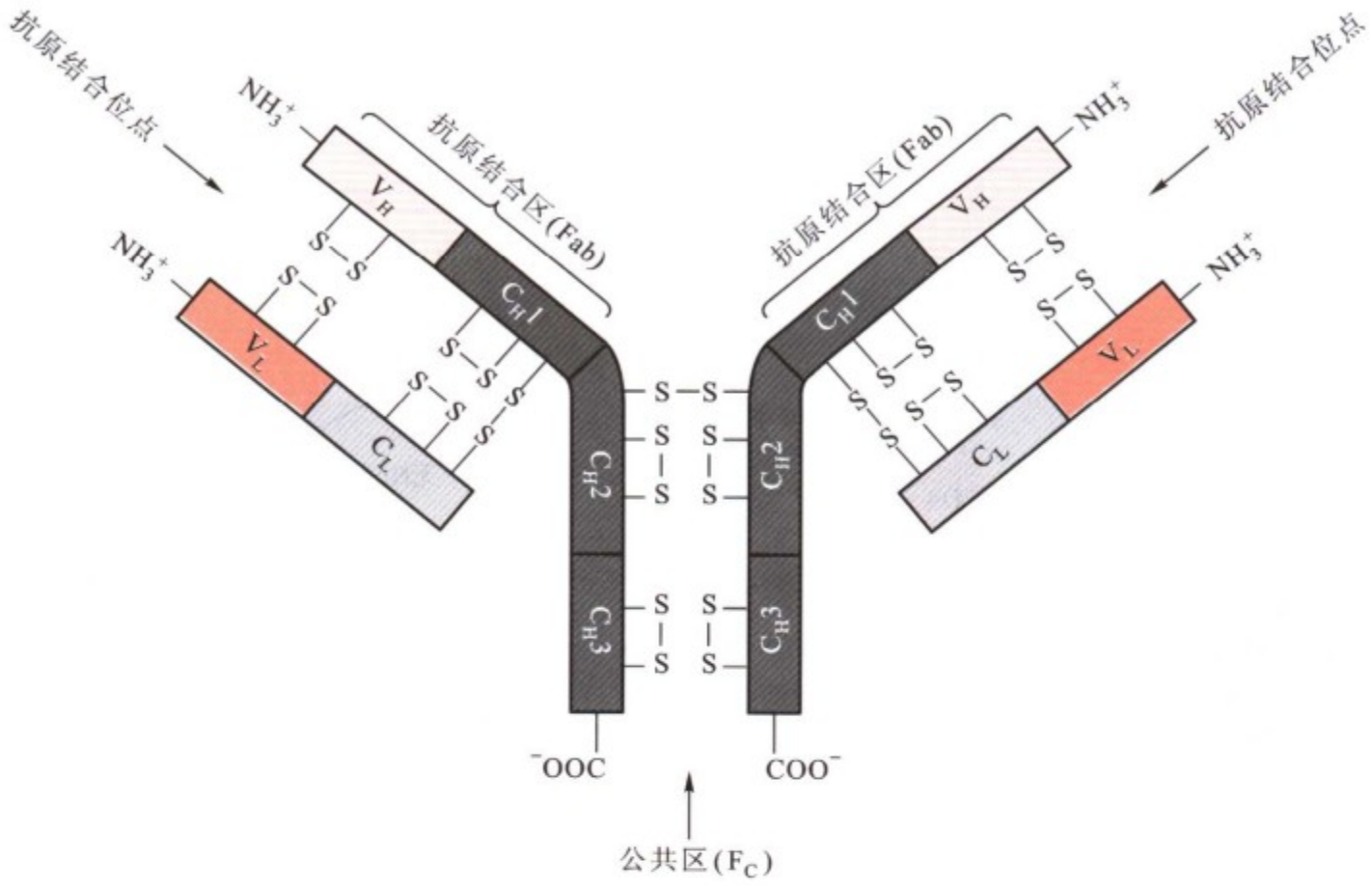



**图 10-6 Ig 分子结构示意图** $V_H$ 重链可变区；$C_H1$，$C_H2$，$C_H3$，重链不变区；$V_L$，轻链可变区；$V_L$，轻链不变区

**表 10-1 抗体的分类**

| 性质 | IgM | IgD | IgG | IgE | IgA |
|---|---|---|---|---|---|
| 重链 | μ | δ | γ | ε | α |
| 结构 | $(U_2L_2)_5J$ | $\delta_2L_2$ | $\gamma_2L_2$ | $\varepsilon_2L_2$ | $\alpha_2L_2$ |
| 占总量 Ig(%) | 5 | 1 | 80 | <1 | 14 |
| 在成年人血清中的平均含量/($g \cdot L^{-1}$) | 1 | 0.03 | 12 | 0.000 3 | 2 |
| 半衰期(d,在血清中) | 5 | 3 | 25 | 2 | 6 |
| 特征 | 出现较早，具有确定补体、有激活巨噬细胞的功能 | 以细胞表面为主，少量存在于血清中 | 激活互补并与巨噬细胞结合 | 刺激肥大细胞，释放组胺 | 分泌蛋白 |

## 10.1.5 免疫球蛋白基因结构

免疫球蛋白(Ig)分子由 *Igk*、*Igl* 和 *Igh* 基因编码，分别位于不同的染色体上。编码一条 Ig 多肽链的基因是种系中数个分隔开的 DNA 片段经重排后产生的。1965 年 Dreyer 和 Bennet 首先提出假说，认为 Ig 的 V 区和 C 区由分隔存在

的基因所编码，在淋巴细胞分化和发育过程中，这些基因片段发生易位重排，形成有转录活性的免疫球蛋白基因。Hozumi 和 Tonegawa 在 1976 年应用 DNA 重组技术证实了这一假说。Tonegawa 因在免疫球蛋白基因结构研究上的重大突破获 1987 年诺贝尔生理医学奖。

分子生物学研究已经阐明，免疫球蛋白基因有 3 个家族，包括两个轻链类型和一个重链类型，这 3 个基因家族位于不同的染色体上(表 10-2)。

表 10-2 免疫球蛋白的基因组成

| 基因家族 | 染色体定位 | | V 基因个数 | | C 基因个数 | |
|---|---|---|---|---|---|---|
| | 人 | 小鼠 | 人 | 小鼠 | 人 | 小鼠 |
| λ | 22 | 16 | <300 | | >6 | 4 |
| κ | 2 | 6 | <300 | ~1 000 | 1 | 1 |
| 重链 | 14 | 12 | ~3 002 | | 9 | 8 |

在生殖细胞基因组中，*V* 基因和 *C* 基因呈分隔独立排列的结构。这种排列类型被称为种系类型(germ line pattern)，存在于所有的体细胞系中。但是，这种排列方式却不存在于有效表达抗体基因的免疫细胞中。研究这类细胞发现，免疫球蛋白基因在表达之前得到了重排。

1. 免疫球蛋白轻链基因结构

(1) κ 基因　位于人类第 2 号染色体上，其可变区和恒定趋分别由不连续的 *V* 基因和 *C* 基因编码。*V* 与 *J* 基因以多拷贝形式存在，其中 *V* 基因簇含有约 300 拷贝。人类和小鼠的 *J* 基因簇都包含 5 个 *J* 基因，其间也有较短的插入序列。*J* 基因簇位于 *V* 基因簇的 3′端，与最后一个 *V* 基因的 3′端相距 23 kb，但与 κ 链恒定区 *C* 基因的 5′端较近，只有约 2.5 kb。人和小鼠的 *C* 基因都是单拷贝，它编码了 κ 链 *C* 区的 106 个氨基酸序列(第 109~214 位氨基酸)。

(2) λ 基因　小鼠只有两个 *Vλ* 基因：*Vλ*1 和 *Vλ*2。人类 *Vλ* 基因簇也位于 2 号染色体上，有 30 拷贝，可分为 *Vλ*1~10 十个亚群，*Vλ*1~5 是主要的亚群，在进化上比较保守。小鼠 *Cλ* 基因簇有 4 个拷贝，而人类 *Cλ* 基因有 6 个以上的拷贝。基因表达过程中，每个 *Vλ* 基因都和 *Cλ* 基因成对排列(图 10-7)。

2. 免疫球蛋白重链基因结构

人类重链基因位于 14 号染色体，由 *V*、*D*、*J* 和 *C* 四种不同基因片段组成。重链基因结构比轻链更为复杂，因此，它在形成抗体多样性方面所起的作用也更大。编码重链 *V* 区基因长约 1 000~2 000 kb，包括 *V*、*D*、*J* 三组基因片段。每个 *V* 基因片段上游有 *L* 基因片段，编码 20~30 个疏水性氨基酸的信号肽(图 10-7)。

重链恒定区基因位于 *J* 基因片段的下游，至少相隔 1.3 kb，它由多个外显子组成。每个外显子编码一个结构域，铰链区(hinge region)由单独的外显子所编码。

$C_H$ 基因中的 Cδ 和 Cα 由 3 个外显子组成，而 Cμ、Cγ 和 Cε 由 4 个外显子组成，分别编码相应重链上的结构域，如 *Cμ* 基因有 4 个外显子编码 μ 链 C 区上的 4 个结构域 Cμ1、Cμ2、Cμ3 和 Cμ4。

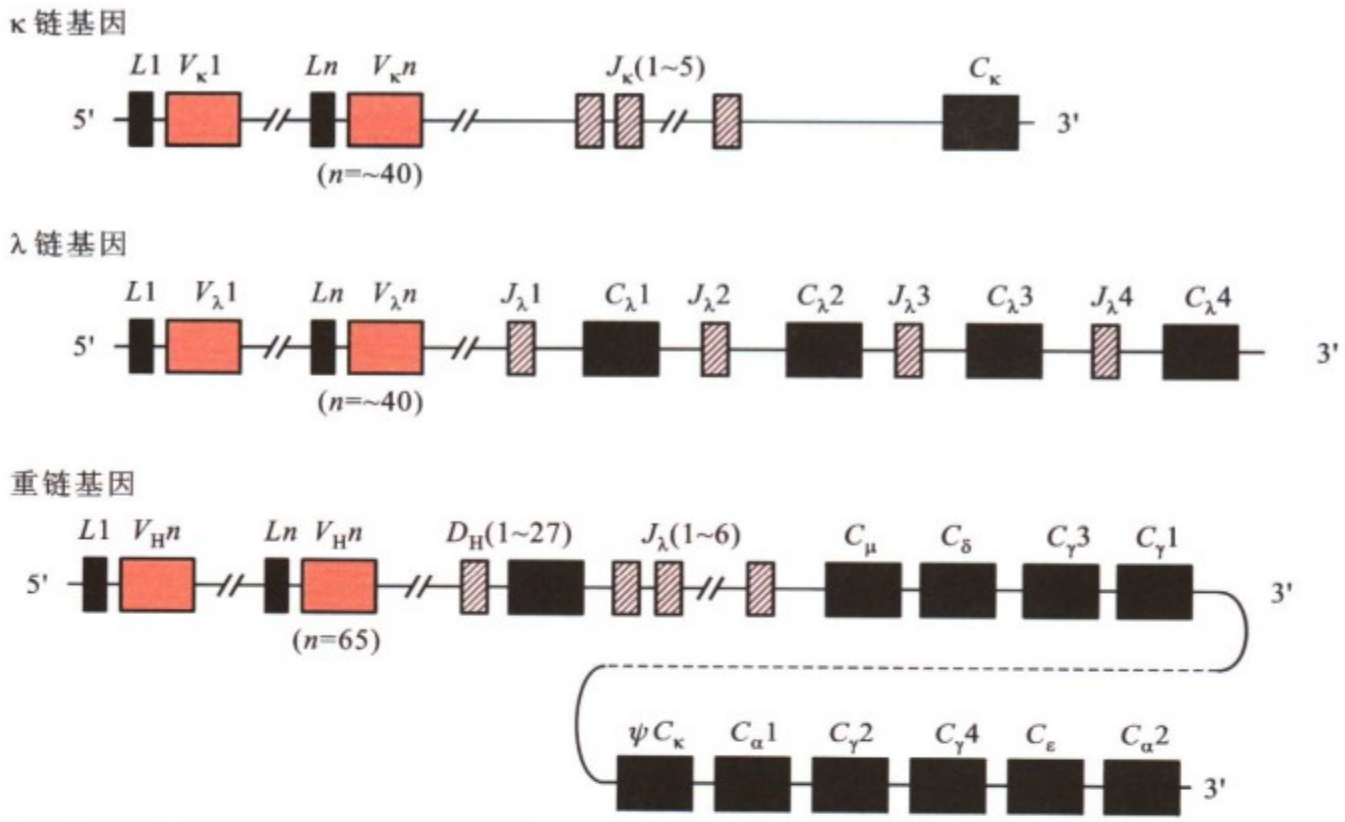



图 10-7 人类 Ig 基因结构示意图

在重链基因中，种系 DNA 上除了有 *V*、*J*、*C* 片段外，在 $V_H$ 和 $J_H$ 之间还有 10 多样性片段(D 片段)，$V_H$ 片段可和 10 个 *D* 片段中任何一个片段重组，*D* 片段则可和 4 个 *J* 片段中的任何一个片段重组。在 *V*-*D* 连接反应和 *D*-*J* 连接反应中，相连接的是同一个 *D* 片段，基因重排后邻近的 $C_H$ 片段得以表达。

### 10.1.6 Ig 基因重排与 DNA 的多样性

一个生物为什么能产生百万种以上($10^6$~$10^8$)的 Ig 分子呢？大多数科学家认为，*V* 区和 *C* 区不同片段在 DNA 分子水平上的各种排列组合是形成 Ig 分子多态的根本原因，这种假设被称为重组种系理论。这就是说，单一种系状态的 *Ig* 基因是不能表达的，只有经过重排，造成种系多样性、连接反应多样性以及体细胞突变等才有可能得到表达。

1. 轻链基因的重排与连接

*V* 基因和 *J* 基因的组合重排是通过 *V* 基因 3′端和 *J* 基因 5′端之间的侧翼序列来实施的。这种侧翼序列又称识别序列，是 *V* 与 *J* 组合重排所不可缺少的。它在结构上的特点是有一段由 7 个碱基组成的反向回文序列——CACAGTG。在 *V* 3′端紧接着回文序列的是 11 个碱基的间隔序列，而后是 9 个富含 A、C 的碱基。*J* 5′端也包含有与上述 7 碱基及 9 碱基区相配对的序列，其间隔序列由 23 个碱基组成，不能与 *V* 3′端 11 个碱基配对，因而形成单链突起(茎环结构)。这种特殊的结构和识别序列有着密切的关系，*V*、*J* 结合后，上述茎环结构被切除，形成连续的 *V*、*J* 片段(图 10-8)。

这种 *VJ* 单位的组合重排是随机的，*V* 基因簇中的任何一个 *V* 片段和 *J* 基因簇中的任何一个 *J* 片段在 DNA 水平上均可随机结合形成 *VJ* 单位。这种随机组合重排又是不精确的，直接造成了重排后的多样性。例如，在重排后，*V*3′和

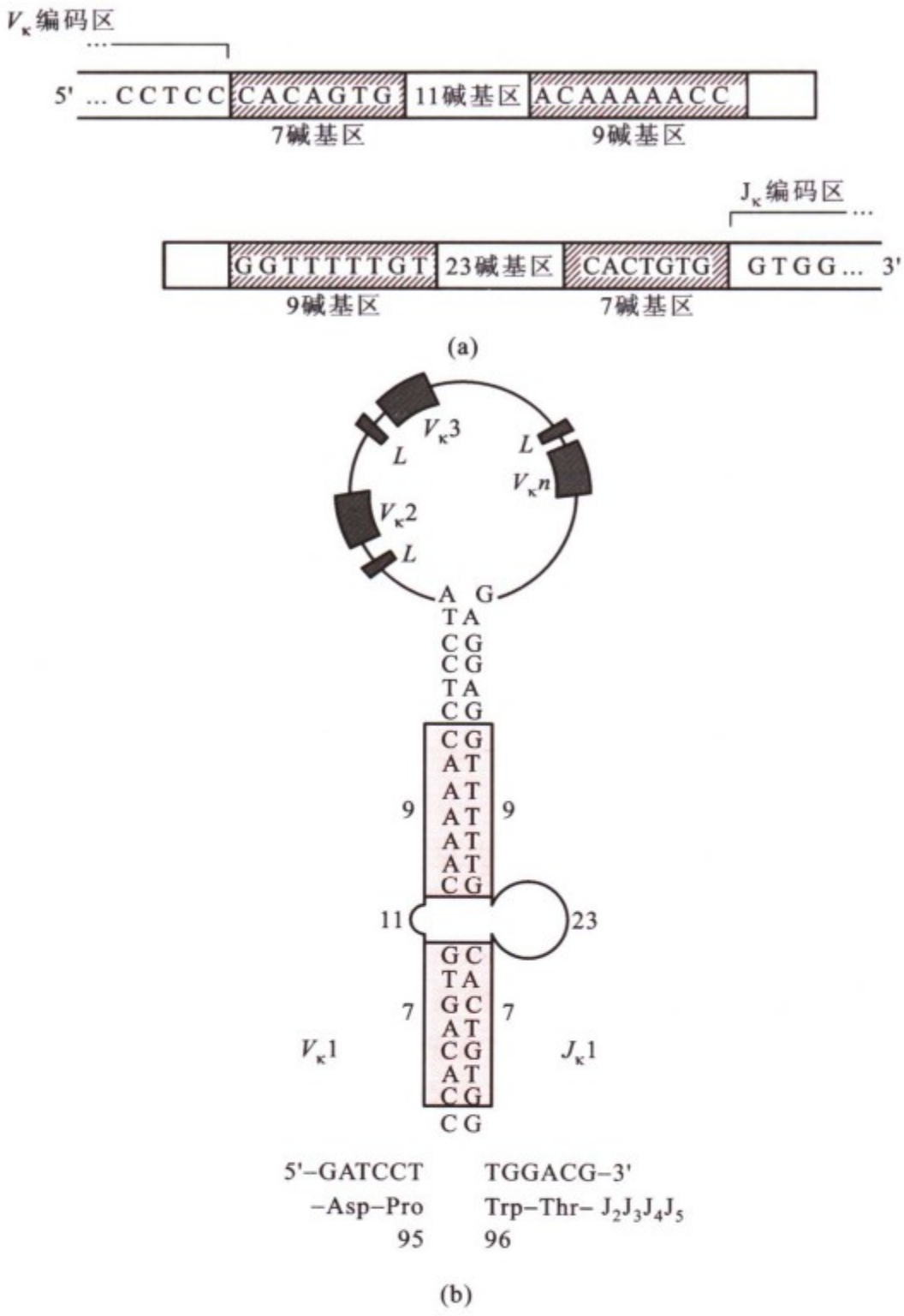


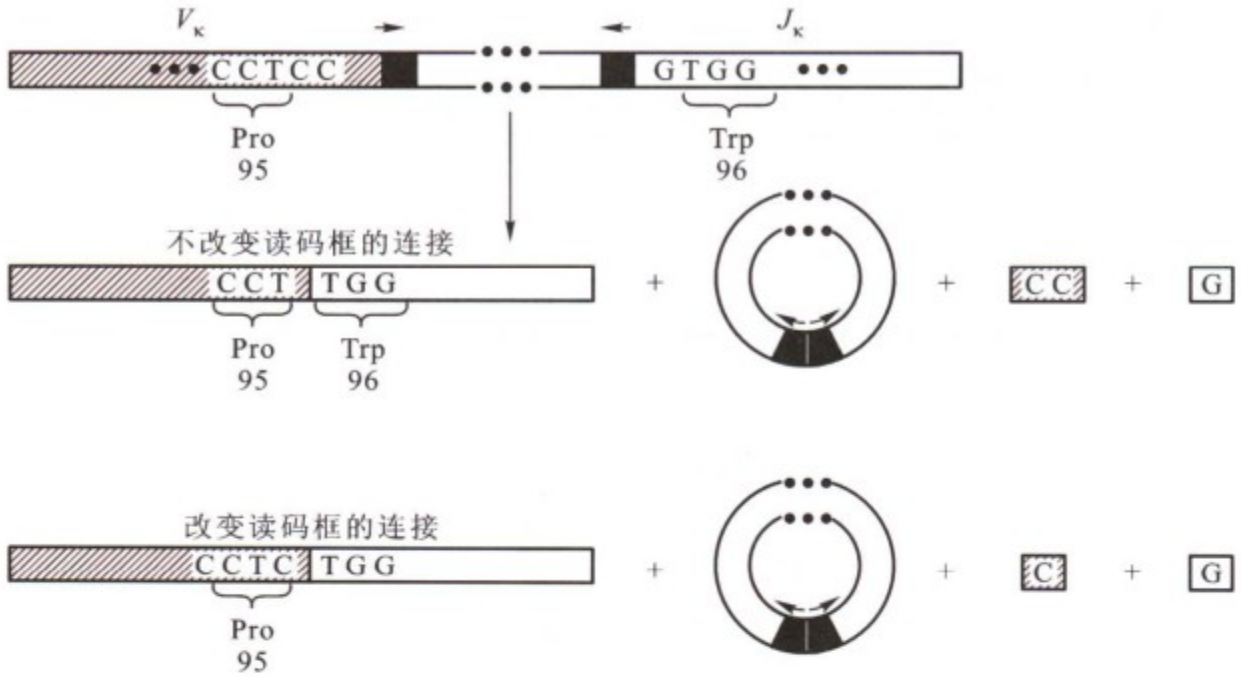

图 10-8 $V_\kappa$与$J_\kappa$基因的重排是通过$V_\kappa$基因3′和$J_\kappa$基因5′端的识别序列完成的

J5′会由于重排不精确而造成碱基丢失或错位，导致κ链中第96位以后的氨基酸编码发生改变，使免疫球蛋白结构发生多样性变化。又如，交互使用*V*和*J*基因间的重组位点，可以产生新的密码子，形成4个不同的重组框架，并在第96位密码子上产生3种不同的氨基酸(图10-9)，其中Arg密码子在种系基因组中是没有的，这是连接反应造成免疫球蛋白多样性的机制之一。

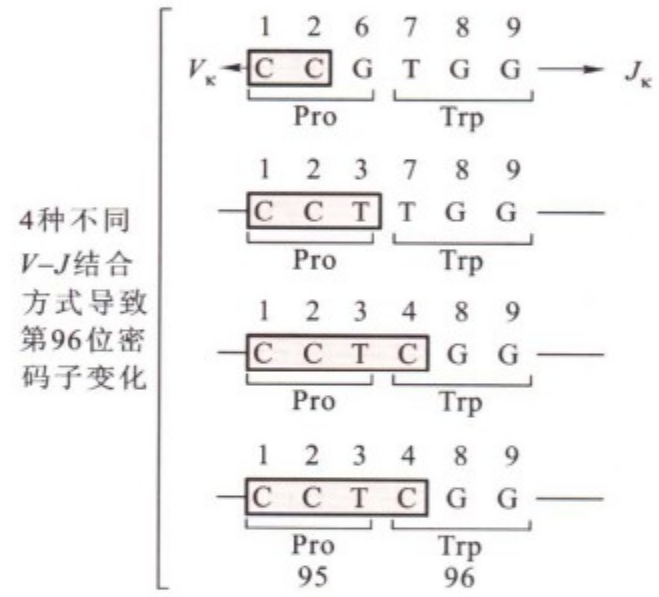
2. 重链基因的重排与连接

重链基因虽然只有一个位点，但结构十分复杂。如小鼠重链基因就包含了大约上百个$V_H$片段、10多个*D*片段和4个*J*片段以及下游17 kb处的$C_H$片段。

*V*基因和*D*及*J*片段组合能产生4 000个不同的*V*区，每个*V*区都能和$C_H$




(a)




(b)

▌图 10–9 *V–J* 连接反应造成 Ig 多样性的分子机制

基因结合。人类的 $C_H$ 基因簇含有 10 个基因(Cμ、Cδ、Cγ3、Cγ1、Cε4、Cα1、Cγ2、Cγ4、Cε、Cα2,其中 Cε4 是假基因)。小鼠 $C_H$ 基因簇只有 8 个基因(依次为 5′–Cμ–Cδ–Cγ3–Cγ1–Cγ2b–Cγ2a–Cε–Cα–3′)。这 8 个 $C_H$ 基因虽然都是有功能的,但在同一细胞中只有一个基因得到表达。

(1) $V_H$、$D_H$ 和 $J_H$ 基因的重排　重链基因组合重排是指在 DNA 水平上由无功能 $V_H$、$D_H$ 和 $J_H$ 基因片段组合重排为有功能 $V_HD_HJ_H$ 基因单位的过程。这个重排过程同样与识别序列有关,而且重链基因要经过两次重排才得以完成。第一次重排发生于 $D_H$ 和 $J_H$ 之间,第二次重排发生于 $V_H$ 和 $D_HJ_H$ 之间,形成有转录功能的 $V_HD_HJ_H$ 基因。第二次重排过程,同样具有不精确性,进一步造成免疫球蛋白的多样性(图 10–10)。

在 $V_HD_HJ_H$ 重排过程中,由于 DNA 聚合酶或脱氧核酸转移酶的作用,有时会在结合处插入一个脱氧核苷酸,形成所谓“N 区”插入序列,致使重链 *V* 区重排

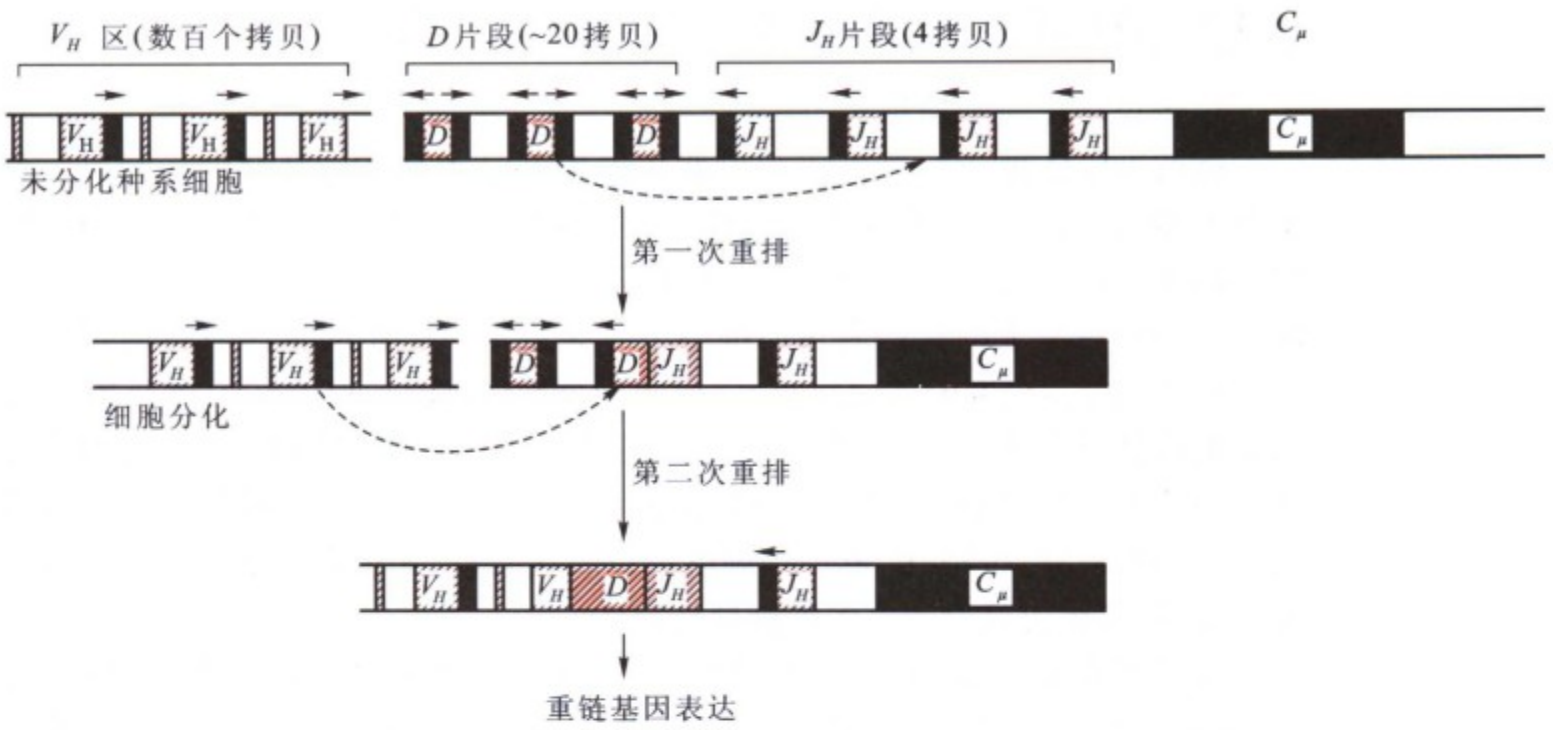



图 10-10 重链基因结构以及伴随细胞分化而来的基因重排

后形成 $V_HND_HNJ_H$ 基因单位。

由于上述重排机制,使重链 $V$ 区更为复杂。如人类基因组有约 300 个 $V_H$ 基因、20 个 $D_H$ 基因和 4 个 $J_H$ 基因及不同的 $N$ 区,加上不同片段之间连接反应的不精确性,$V_HD_HJ_H$ 重排结合产生的变化达到 300×20×4×100(N 区因素)×10(连接不精确因素)=$2.4\times10^7$。因为 $V_\kappa$ 和 $J_\kappa$ 重排结合的多样性程度是 $5\times10^3$,因此,由κ链和重链组成的 Ig 分子抗体的多样性就可能有 $2.4\times10^7\times5\times10^3=1.2\times10^{11}$ 之多。

(2) $V_HD_HJ_H$ 基因单位与 $C_\mu$ 基因或 $C_\delta$ 基因在 RNA 水平上的剪接。$V_HD_HJ_H$ 基因簇作为一个转录单位先转录成 mRNA 前体,经剪接加工形成 μm、μs 和 δ3 种 mRNA。具体形成哪一种 mRNA,则要根据 $C_\mu$ 基因中加 poly A 的部位而定。$C_\mu$ 基因区有两个部位可加入 poly A,一个是在 $\mu m$ 和 $\mu s$ 之间,一个是在 $\mu m$ 下游区。如果 poly A 加到 $\mu m$ 与 $\mu s$ 之间,经剪接形成 $\mu s$ mRNA;如果 poly A 加到 $\mu m$ 下游区,经剪接后形成 $\mu m$ mRNA。$Ms$ mRNA 经剪接加工形成分泌型 $C_\mu$ mRNA,$\mu m$ mRNA 则形成膜结合型 $C_\mu$ mRNA。由于 $C_\mu$ 基因和 $C_\delta$ 基因间的距离很短,因此,$V_HD_HJ_H$ 在 RNA 水平上既可与 $C_\mu$ 结合也可与 $C_\delta$ 结合。在淋巴细胞中,$V_HD_HJ_H$ 和 $C_\mu$ 结合重排产生 IgM,和 $C_\delta$ 结合重排产生 IgD。

(3) 体细胞突变引起的多样性 虽然可变区的多样性主要是由于 DNA 重排和连接反应所引起的,体细胞的突变也是重要原因之一。研究发现,$V$-$J$ 或 $V$-$D$-$J$ 结合后,体细胞中可变区的突变频率极高,比一般高出 $10^4$~$10^5$ 倍。突变发生的部位大都在 $V$-$D$-$J$ 外显子内或其近旁,恒定区内突变率则较低。体细胞突变是免疫系统以有限的遗传信息生成海量抗体分子以应付外来抗原的一种极为精密的“微调”,起到了专一性修饰免疫球蛋白的作用。体细胞突变的原因可能是受未知“诱发剂”影响的自我驱动或抗原驱动,也有人认为可能存在着特殊的高突变机制,这有待于进一步研究。

### 10.1.7 免疫球蛋白基因表达

1. 等位排斥与同型排斥

所谓等位排斥是一个 B-淋巴细胞中发生 *V-J* 和 *V-D-J* 重排时，只有一种重组类型出现，产生一种轻链和一种重链。淋巴细胞中产生免疫球蛋白的基因位于两条同源染色体上，而免疫球蛋白基因的表达只发生在其中一条上，这种因为一条染色体上的基因表达而抑制另一条染色体上相同基因表达的现象称为等位基因排斥(allelic exclusion)。而同型排斥(isotypic exclusion)则是指B-淋巴细胞的轻链表达时，只生成一种链(κ 链或是 λ 链)，不可能同时表达 κ 链和 λ 链。等位排斥和同型排斥使 B-淋巴细胞无性系只表达一种特异抗体的事实得到了很好的解释。否则，B-淋巴细胞既可以表达 κ 链，又可以表达 λ 链上的两个等位基因，细胞内就会产生多种特异性抗体。

为了区分免疫球蛋白基因的各种状态，科学上将未发生重排的种系构型等位基因称为 $Ig^{\circ}$，将发生重排并进行功能表达的等位基因称为 $Ig^{+}$，将发生无效重排的无活性等位基因称为 $Ig^{-}$。细胞中有可能存在下述几种情况：①等位基因之一发生有效重排，产生活性基因，并产生等位排斥，使另一基因的重排和表达受到抑制，这个细胞构型为 $Ig^{\circ}/Ig^{+}$；②等位基因之一发生无效重排，另一个染色体上的基因仍处于种系构型，则该细胞构型为 $Ig^{\circ}/Ig^{-}$，这样的细胞构型不会抑制另一个等位基因的重排；③第一次无效重排后，细胞内又发生另一次有效重排，生成 $Ig^{+}/Ig^{-}$，阻止了进一步重排的发生。

2. 增强子与 Ig 基因表达

有研究发现，在免疫球蛋白 *V* 区基因的转录过程中，浆细胞 IgA 的转录比 B-淋巴细胞中快 4 个数量级。那么，是什么因素在起作用呢？一般认为，在 *Ig* 基因的转录过程中，除了启动子的正常功能以外，增强子的作用不可忽视。

如图 10-11 所示，对免疫球蛋白重链基因来说，DNA 重排以后，靠近 3′ 端的间隔序列被切除，而 5′启动区无任何变化。但为什么该启动子在基因重排之前不支持转录，而重排以后获得了转录功能呢？有科学家将该免疫球蛋白基因用体外重组技术导入肉瘤细胞系并进行缺失突变，发现缺失 *J-μ* 之间的 DNA

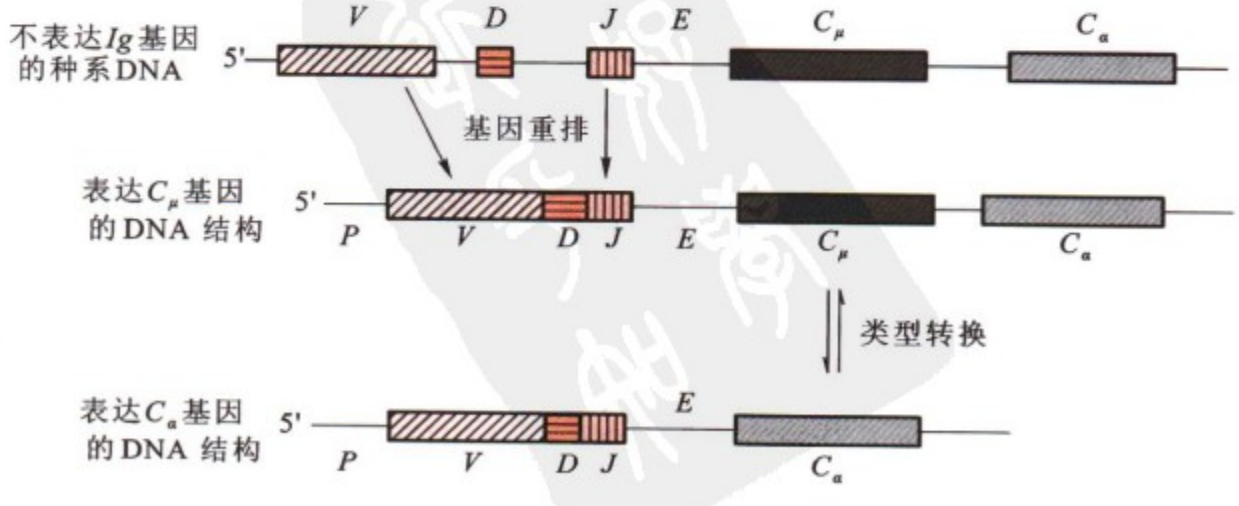



图 10-11 增强子序列(E)对免疫蛋白重链基因表达的影响

序列导致基因失去表达活性;若将这段序列重新导入该重组子内原有位置或附近位点,不论方向如何,都能支持免疫球蛋白基因的表达,证实可能有增强子存在于该 DNA 序列内。*V* 基因的启动子只有经过重排、缺失部分 DNA 序列以后,才受这一增强子的影响,获得高效表达。进一步的实验还证实,这个增强子同时具有组织特异性,随着 B-淋巴细胞成熟为浆细胞,它对 *Ig* 基因表达的影响越来越大。

### 10.1.8 主要组织相容复合体的表达调控

1. 主要组织相容性复合体概述

引起移植排斥的细胞表面抗原称为组织相容性复合体(major histocompatibility complex,MHC)。由于这些抗原的存在,同一物种某个体的组织移植至另一个在抗原组成上不相容的个体时,就会引起免疫反应,产生对所移植器官的排斥。按其诱导排斥的急缓强弱而有主要和次要之分,其中 MHC 在移植排斥中起决定作用。一般认为,MHC 不仅控制着种内移植排斥反应,而且与机体免疫应答、免疫调节及某些病理状态的产生密切相关,处于机体特异免疫反应的中心地位。一方面,细胞毒性 T 细胞通过其表面的 TCR 对呈递于 MHC 分子表面的异己外源肽段进行识别,并直接杀伤呈递这类 MHC 的细胞;另一方面,辅助 T 细胞也通过其外表面 TCR 识别Ⅱ型 MHC 蛋白中的外源肽段,辅助 B 细胞产生抗体(图 10-12)。所以,MHC 与 TCR、Ig 相互配合,构成了生物体内复杂多变的免疫反应和应答。

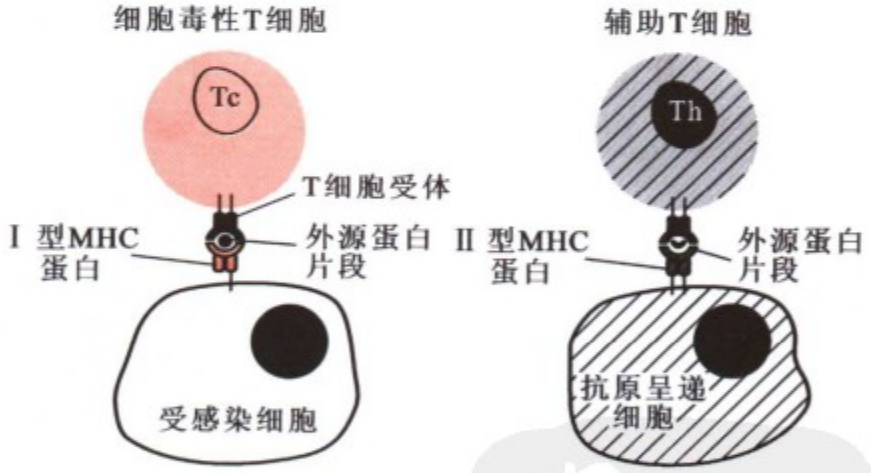



**图 10-12 细胞毒性 T 细胞和辅助 T 细胞与不同的 MHC 分子相互作用发挥其生物学功能**
细胞毒性 T 细胞能识别所有与Ⅰ型 MHC 分子相结合的受感染细胞携带的外源蛋白片段,辅助 T 细胞只能识别抗原呈递细胞(如巨噬细胞和 B 细胞)表面所携带的外源蛋白。与Ⅰ型 MHC 分子相结合的抗原主要是由受感染细胞产生的,而与Ⅱ型 MHC 分子相结合的抗原主要是由抗原呈递细胞通过胞饮现象吸收、加工以后呈递细胞表面的。在组织移植时,细胞毒性 T 细胞主要对外源Ⅰ型 MHC 做出反应,而辅助 T 细胞主要对外源Ⅱ型 MHC 做出反应。

2. MHC 的分类

早期有关 MHC 的研究主要集中于小鼠,它的主要组织相容性复合体称为 H-2(histocompatibility antigen-2),该基因簇位于第 17 号染色体上。人 MHC 又

称白细胞抗原(human leukocytoantigen,HLA)体系,是到目前为止免疫反应中最复杂,最具多态性的体系。基因簇位于第 6 号染色体,全长约 2 000~3 500 kb。HLA 共分为 4 型:Ⅰ型分子包括 HLA-A、HLA-B 和 HLA-C,广泛存在于各种组织细胞中;Ⅱ型分子包括 HLA-DP、HLA-DQ 和 HLA-DR,存在于 B 细胞、巨噬细胞和活化 T 细胞中;Ⅲ型分子为补体系统,包括 C2 和 C4 位点,存在于血清中;Ⅳ型分子可能是一些分化抗原,只存在于淋巴细胞、某些细胞毒性 T 细胞和白细胞中。

Ⅰ型分子有一个信号肽链,称为移植抗原(transplantation antigen),其相对分子质量约为 $4.5\times10^4$,是跨膜蛋白,与免疫球蛋白结构域相似,可被分为 3 个部分。在细胞质内的部分是较短的亲水 C 端片段,称为胞质片段,横跨脂质双层的一个较短片段称为跨膜片段,在细胞膜外侧是个较大的 N 端片段,称为胞外片段。该片段可进一步分成 3 个结构域:$\alpha_1$、$\alpha_2$ 和 $\alpha_3$,前两个结构域可变度高,呈多态性。它们在细胞表面与一种由另一条染色体上的基因所编码的,被称为 $\beta_2$-微球蛋白的较小蛋白质共价相连(图 10-13)。

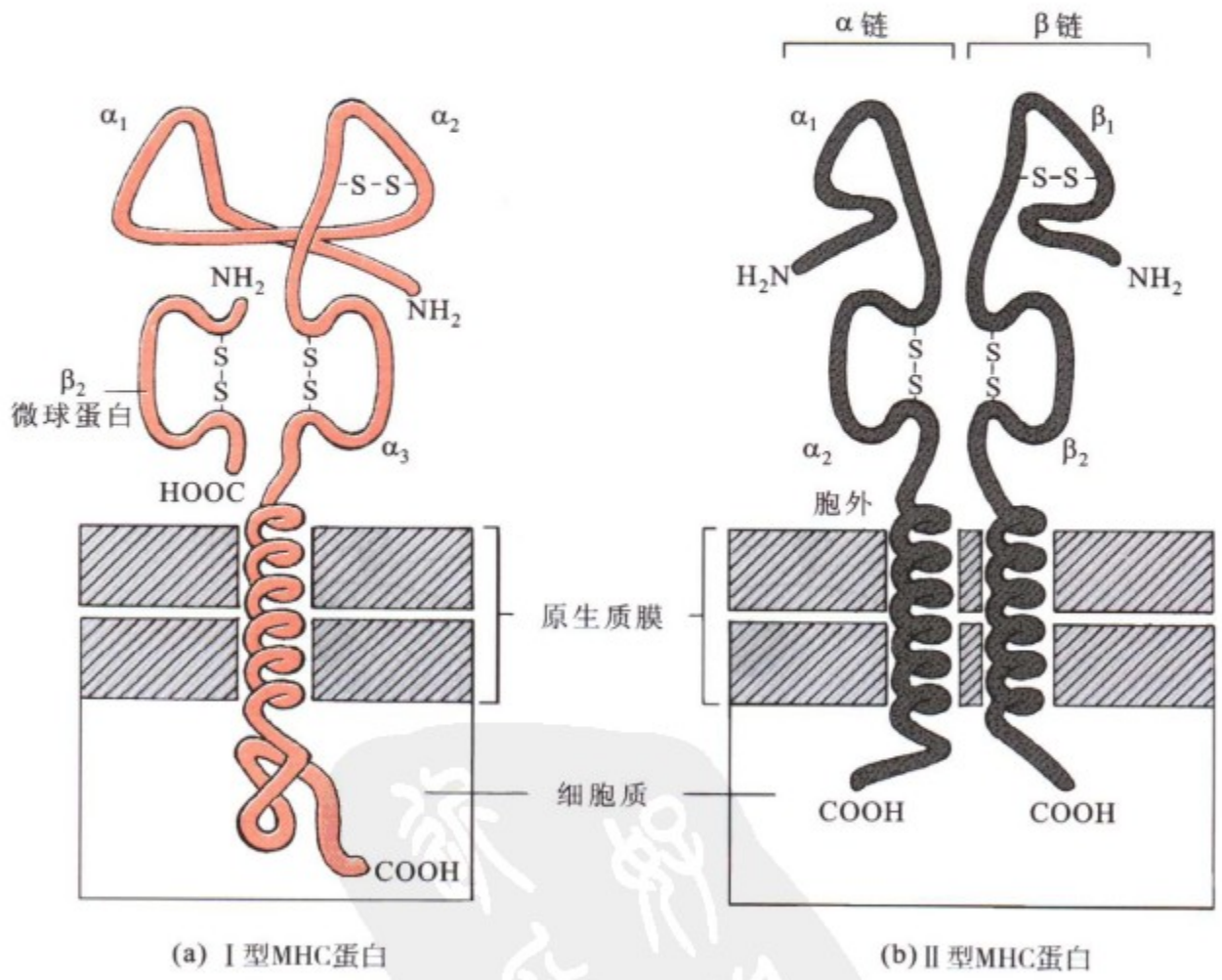



图 10-13 Ⅰ型和Ⅱ型 MHC 的组成和结构域分析

Ⅱ型分子是 HLA 系统中的另一个重要成分,由 α 和 β 两条糖化多肽链组成。α 链为 $3.3\times10^4$,β 链为 $2.8\times10^4$。β 链上携带有主要多态决定簇。已经发现有 6 个 α 基因和 7 个 β 基因。在 6 个 α 基因中,两个编码 DP,两个编码 DQ,有一个编码 DR,还有一个编码未知抗原。在 7 个 β 基因中,两个编码 DP,两个编码

DQ,3 个编码 DR。HLA 区域涉及大约 $3.0\times10^6$ bp 序列,足以编码 150 个以上相对分子质量在 $4.0\times10^4$ 左右的蛋白质，现已克隆了 30 多个基因和假基因。表 10-3 比较了Ⅰ型和Ⅱ型 MHC 的主要性能。

表 10-3 Ⅰ型和Ⅱ型 MHC 主要性能比较

| | Ⅰ 型 | Ⅱ 型 |
|---|---|---|
| 蛋白质的结构 | α 链 + $\beta_2$ - 微球蛋白 | α 链 + β 链 |
| 在细胞内分布 | 细胞毒性 T 细胞 | 辅助 T 细胞 |
| 抗原蛋白来源 | 细胞质蛋白 | 内吞膜蛋白和细胞外蛋白 |
| 多态性区域 | $\alpha_1 + \alpha_2$ | $\alpha_1 + \beta_1$ |

3. MHC 的分布及主要作用

MHC 的组织特异性表达是受发育调控的。早期胚胎细胞不表达 MHC 抗原,直到发育至体节中期才开始表达。在成熟体细胞中,Ⅰ型 MHC 的分布极为广泛,以淋巴样细胞中表达量最高,而在脑组织和精细胞中几乎不表达。Ⅱ型 MHC 的分布范围相对说来比较狭窄,主要存在于一些免疫细胞,如 B 细胞、巨噬细胞、单核细胞的表面,在树突细胞和内皮细胞中也有分布。

抗原呈递是 MHC 分子的重要功能。研究证实，当细胞被外源微生物感染时，细胞毒性 T 细胞可通过 T 细胞受体与带有外源蛋白的Ⅰ型 MHC 分子相互作用，导致这些细胞的最终降解。存在于 B 细胞及巨噬细胞内的Ⅱ型 MHC 分子,捕捉入侵抗原以后,主要与辅助 T 细胞相互作用,激活 B 细胞的 Ig 系统或诱导巨噬细胞消灭入侵微生物(图 10-12)。Ⅰ型、Ⅱ型 MHC 在组织分布上的差异可能与它们呈递不同的抗原有关。Ⅰ型主要负责胞内抗原的呈递。胞内蛋白质被加工成抗原性肽段后被转运至内质网腔内,与Ⅰ型分子形成抗原复合物,经高尔基体分泌的小泡运输到质膜上呈递给 Tc 细胞,激活细胞免疫过程。Ⅱ型MHC 则主要负责胞外抗原的呈递。内质网中Ⅱ型 MHC 的 α、β 链只有在结合形成异源二体时才具有相对稳定性,再通过不同途径与胞外抗原结合,转运到质膜表面,呈递给 Th,激活体液免疫过程。

一般情况下,MHC 分子与 TCR 的相互作用往往不足以维持这两类细胞间的牢固联系，需要有另一类称为附加受体(accessory receptor)的分子协同作用。当这类附加受体也具有利用自身胞间信息系统激活 T 细胞的功能时,就被称为共受体(co-receptor)。这类受体不同于 TCR 或 MHC 分子,它们不能单独与抗原结合,没有可变区,也不具有多型性。CD8 和 CD4 共受体分别参与Ⅰ型 MHC 与细胞毒性 T 细胞及Ⅱ型 MHC 与辅助 T 细胞之间的反应(图 10-14)。

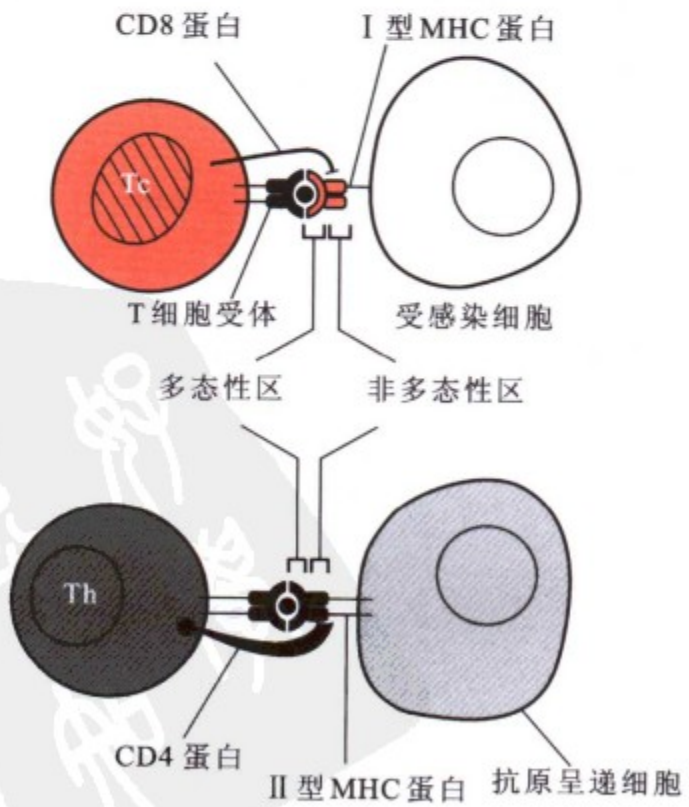



图 10-14 CD4 和 CD8 共受体参与辅助 T 细胞和细胞毒性 T 细胞介导的免疫反应

## 10.2 果蝇的发育与调控

果蝇是一种双翅目昆虫，其幼虫和成虫依赖于正在腐烂的果实。果蝇的个体小、生命周期快、繁殖容易，每只雌虫两周就可以产生约 300 个后代。此外，果蝇的巨大多线染色体特性使它最适合于遗传分析和基因定位，是基因与发育领域最好的模式生物之一。Thomas Morgan 因为描述了果蝇的遗传重组原理及发现了性染色体，于 1933 获得诺贝尔生理医学奖。Morgan 的同事H.J. Mülter 因为发现 X 线诱导突变而于 1964 年获得诺贝尔奖。

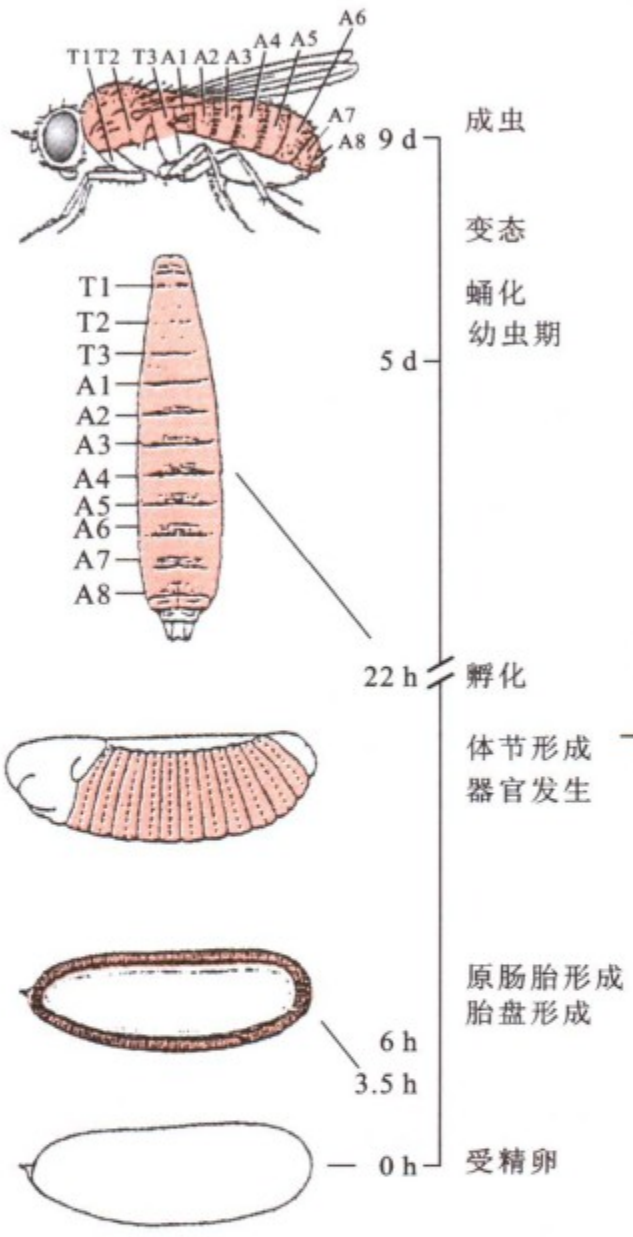



图 10–15 果蝇发育的不同时期分析

果蝇发育的时间周期取决于温度，在 25℃下约 9~14 d 为一个周期(图 10–15)，其中第一天为胚胎发育期。幼虫经历三个阶段，到第四天蜕皮分化后蛹化(popation)，在蛹中经过 5 d 的变态，再发育为成虫。另外，还应加上受精所需的1~2 d，共约两周左右。成年果蝇存活约 9 d。果蝇发育的许多特性反映出这非同一般的快速生命周期在进化过程中的演变。

### 10.2.1 卵子发育

果蝇卵细胞在产生过程中就已经开始快速的胚胎发育。由于这个特性，果蝇的许多遗传发育特点受到高度关注。果蝇卵在卵巢管中形成，卵巢管被横向的管壁分成许多小室(图 10–16)。构成卵巢管管壁并围绕未来卵细胞的细胞叫滤泡细胞(follicle cells)，类似于脊椎动物卵巢中的卵泡细胞。每个小室中含有一个雌性原始生殖细胞，即卵原细胞(oogonium)。

卵原细胞经过四次有丝分裂成为 16 个细胞，称为并合体(fusomes)的这些细胞通过胞质桥构成内部连接，其中之一以后将发育成为卵母细胞(oocyte)，其余 15 个姐妹细胞将发育成为抚育细胞(nurse cell)。卵母细胞为双倍体，完成减数分裂后成为单倍体。抚育细胞由于反复进行 DNA 复制而成为多倍体，这种基因组的前期扩增为在以后的胚胎发育中保持很高的转录活性提供了物质保障，并为卵母细胞提供大量的核糖体和包裹在核糖体蛋白颗粒中的 mRNA。这些核糖体和 RNP 由滋养细胞排出后经并合体运输进入卵母细胞，供正在发育的卵使用。

滤泡细胞有自己的祖先。在卵子发生后期，滤泡细胞为卵母细胞提供营养，尤其是保证卵黄的供应。这些卵黄首先在卵细胞外的“脂肪体”中以卵黄原蛋白和卵黄高磷脂蛋白的形式合成并被释放到血淋巴中，被滤泡细胞收集后再传递给卵母细胞，储存于卵黄板和卵黄颗粒中，为胚胎发育提供氨基酸、磷酸和能

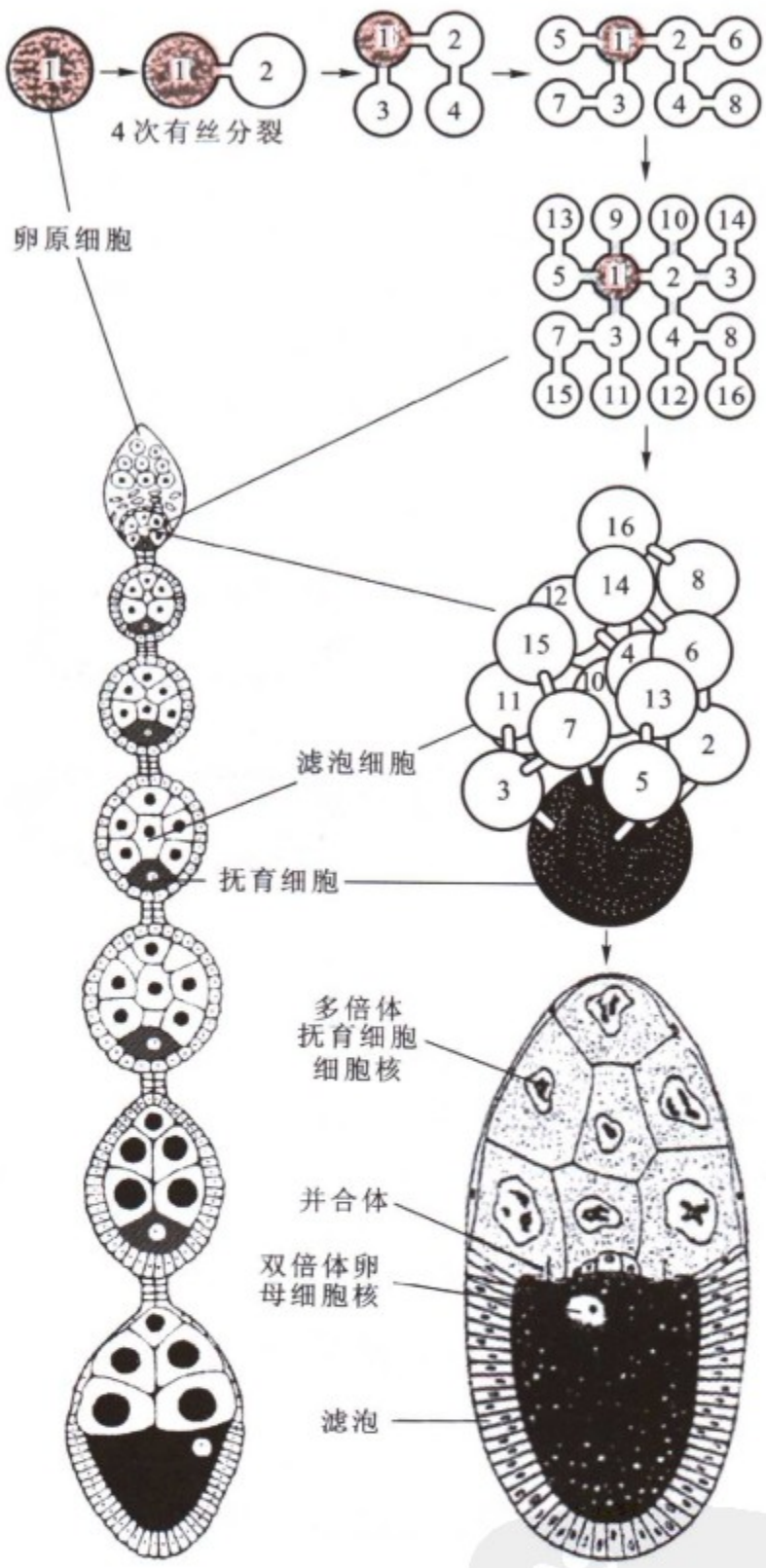



图 10-16 果蝇卵子形成过程示意图

量。滤泡细胞还为胚胎发育提供末端和腹背分化的信号。

卵母细胞自身的细胞核不具转录活性,而由母源抚育细胞、滤泡细胞和脂肪体细胞利用自身的基因和细胞资源提供遗传信息和营养物质,然后输入到卵母细胞中。这些被输入到卵母细胞的基因被称为母源影响基因(maternal effect genes),对卵子的发育有重大影响。

## 10.2.2 胚胎发育

1. 胚胎发生的准备——卵裂

果蝇的胚胎发育得非常快(图 10–17),在产卵后立即开始,并在 1 d 内孵化成幼虫。果蝇的卵裂是异常的。首先,细胞核以每 9 min 一次的高频率复制,直至达到约有 6 000 个核出现为止。在这段时间,卵子发育成为一个合胞体(syncytium)。当出现 256 个以上的细胞核时,这些细胞核开始向卵的外周移动,并定居到皮层组织。此时,细胞质膜沿核间内陷,细胞质环绕每个核封闭成一个个小室(细胞产生),这就是细胞胚盘期。胚盘的腹部构成生殖带,产生胚胎固着层。在果蝇的细胞胚盘产生以后,其身体的发育和形成从生殖带腹部开始,并发展到卵背侧,整个发育过程涉及中胚层的形成、神经索和脑的形成等过程。

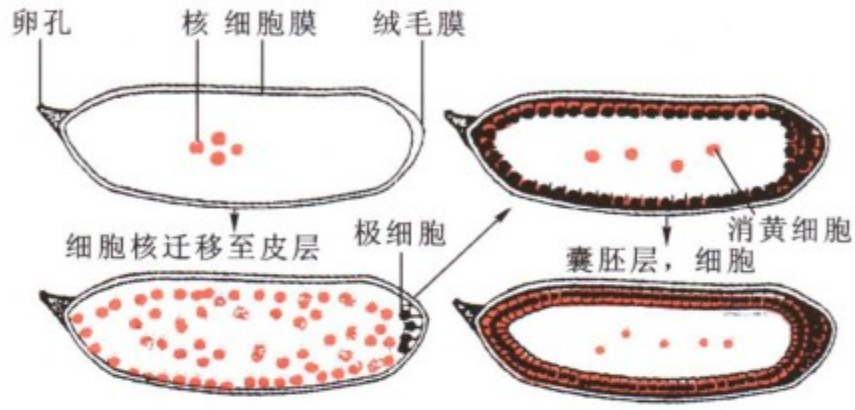



图 10–17 果蝇的卵裂过程简图

2. 果蝇胚胎的极性与躯体的基本模式

与其他两侧对称的动物相似,果蝇的卵、胚胎、幼虫和成体都具有明确的前–后轴和背–腹轴。果蝇形体模式的形成是沿前–后轴和背–腹轴进行的。果蝇胚胎和幼虫均沿前–后轴显示规律性分节,分属 3 个解剖区,从前到后被称为头节,3 个胸节及 8 个腹节。果蝇幼虫的前后端又特化产生原头和尾节(图 10–18)。

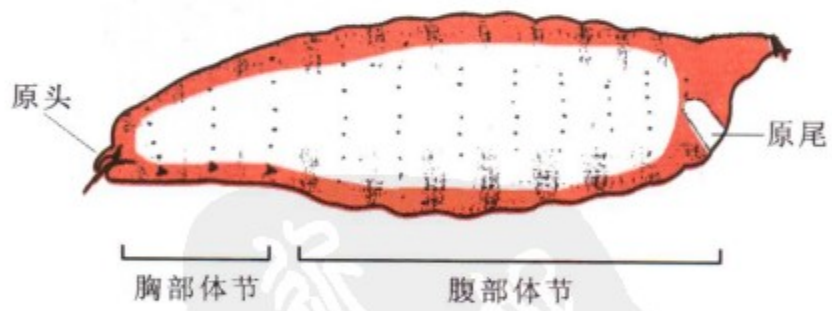



图 10–18 果蝇 1 龄幼虫形体结构模式

在果蝇中,轴决定是在母源基因而非胚胎本身基因的调控下发生的。这些轴决定基因在母体输卵管中的抚育细胞或滤泡细胞中很活跃,以核蛋白形式将 mRNA 运输到卵子中发挥作用。已发现有 4 组母源影响基因与果蝇胚轴的形成有关,其中 3 组参与胚胎前–后轴,即前端系统决定头和胸部的分节,后端系统决定腹部的分节,末端系统决定胚胎两端不分节的原头区和尾节,第四组基因决定胚胎的背–腹轴。

早在 20 世纪初,胚胎学家就已经注意到很多动物受精卵中特定部位的细

胞质定位与胚胎发育直接相关。如果使果蝇卵前端少量的细胞质流失，该卵发育成的胚胎就会缺失头部和胸部。这种缺失突变体与后来发现的 bicoid 突变体的表型相似。卵后端的极质流失时，胚胎缺失腹部而并不影响最后的尾节。卵子其他部位细胞质的少量流失不会影响整个身体模式形成，说明卵子前、后端的细胞质中含有与决定果蝇身体模式形成有关的信息物质存在。

在果蝇发育的卵细胞阶段，由于抚育细胞和卵泡细胞的分泌和诱导，在长椭圆形的卵细胞两端分别差异性地锚定和贮存了特异性翻译调控因子 BICOID 和 *nanos* mRNA。进入胚胎发育阶段后，这两种信使 RNA 翻译成蛋白质，并在胚胎中形成了 BICOID 由前向后梯度性递减和 NANOS 梯度性递增分布。由于 BICOID 蛋白对 *caudal* mRNA、NANOS 蛋白对 *hunchback* mRNA 的翻译分别具有抑制作用，使这两种原本均匀存在于胚胎中的 mRNA 产生了蛋白表达的梯度差异，即 HUNCHBACK 蛋白含量由前向后逐渐降低，而 CAUDAL 蛋白含量则渐次增高(图 10–19)。

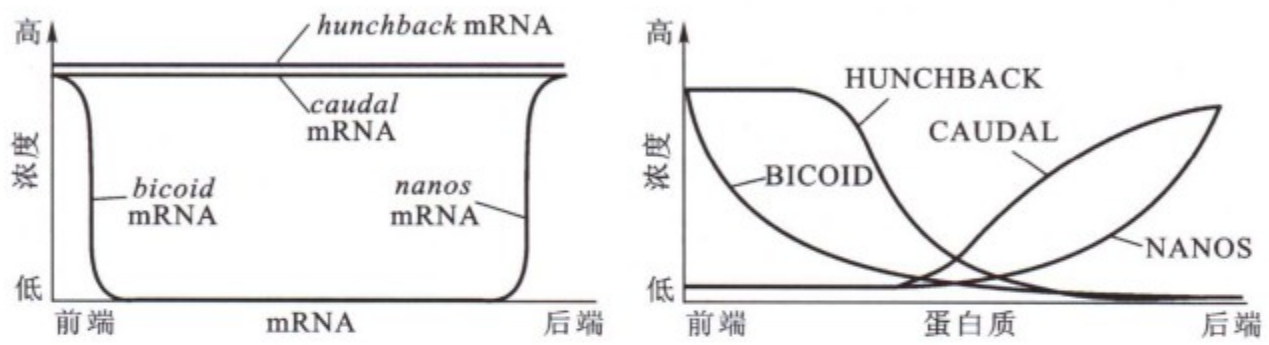



图 10–19 四种母源影响基因的 mRNA 和蛋白沿果蝇卵子和胚胎前–后轴分布的浓度变化图

果蝇胚胎、幼虫和成虫的前后极性均源于卵子期发生的极性。有 4 个非常重要的形态发生决定基因参与调控胚胎前–后轴形成，*biocoid* 和 *hunchback* 调控胚胎前端结构的形成，*nanos* 和 *caudal* 调控胚胎后端结构的形成。上述形态发生决定基因表达以后，首先激活间隙基因表达，再由间隙基因激活成对控制基因，由成对控制基因激活体节极性基因表达。同时，间隙基因、成对控制基因及体节极性基因产物与同源域基因(homeotic gene)上游调控区发生相互作用，调节同源域基因表达，最终决定了每个体节的命运(图 10–20)。

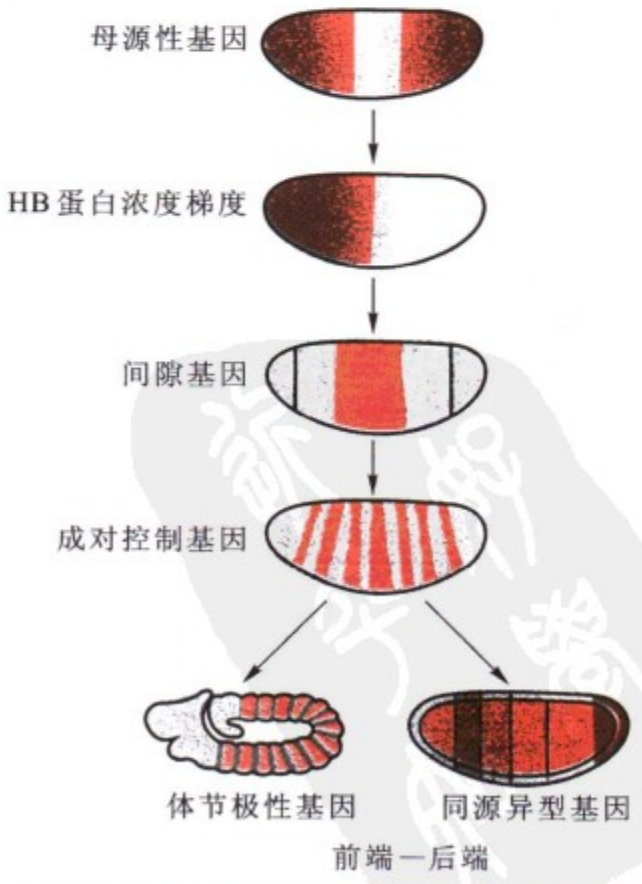



图 10–20 果蝇胚胎模式建立过程中关键基因的表达顺序

### 10.2.3 果蝇的末端系统及背腹极性基因与发育调控

如果控制前端和后端系统的基因都发生突变，果蝇胚胎仍可产生某些前后模式，并发育成具有两个尾节的胚胎，暗示还存在第三个前–后轴确定系统，即末端系统。末端系统包括约 9 个母源效应基因。这些基因缺失会导致胚胎的前端原头区和后端尾节（不分节部分）缺失。在这个系统中发挥关键作用的是 *torso* 基因。该基因编码一种跨膜酪氨酸激酶受体，其 N 端序列位于细胞膜外，C 端位于膜内。在卵子发育中，*tor* 基因在整个合胞体胚胎的表面表达（图 10–21）。受精后配体被释放出来并穿过卵黄膜进入围卵隙。只有当胚胎前、后、末端细胞外存在配体（信号分子）时，才能使 *tor* 特异性活化，导致胚胎前、后、末端细胞特化。

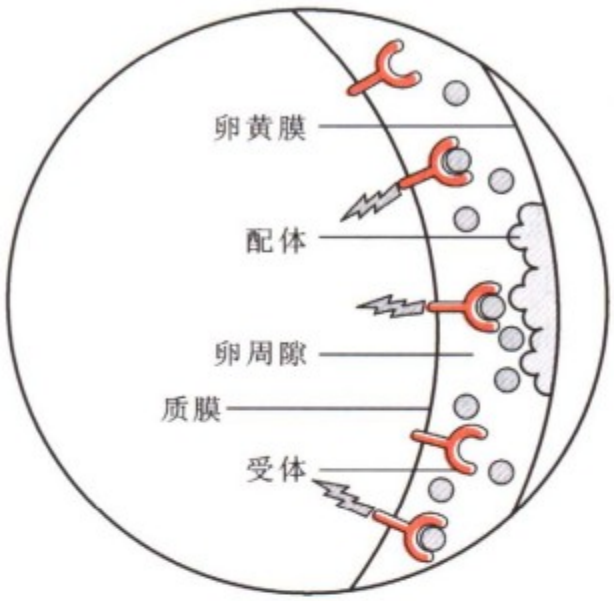



图 10–21 *tor* 基因在控制果蝇胚胎末端的分化中的作用

母源性 *toll* mRNA 产物是卵细胞的跨膜受体，其作用是感知外部信号并提示胚胎在何处产生腹侧（图 10–22）。*toll* 受体与调控果蝇腹侧发育的外部信号可能是由一种蛋白酶从锚定复合体中释放出并定位于卵细胞周围外卵黄

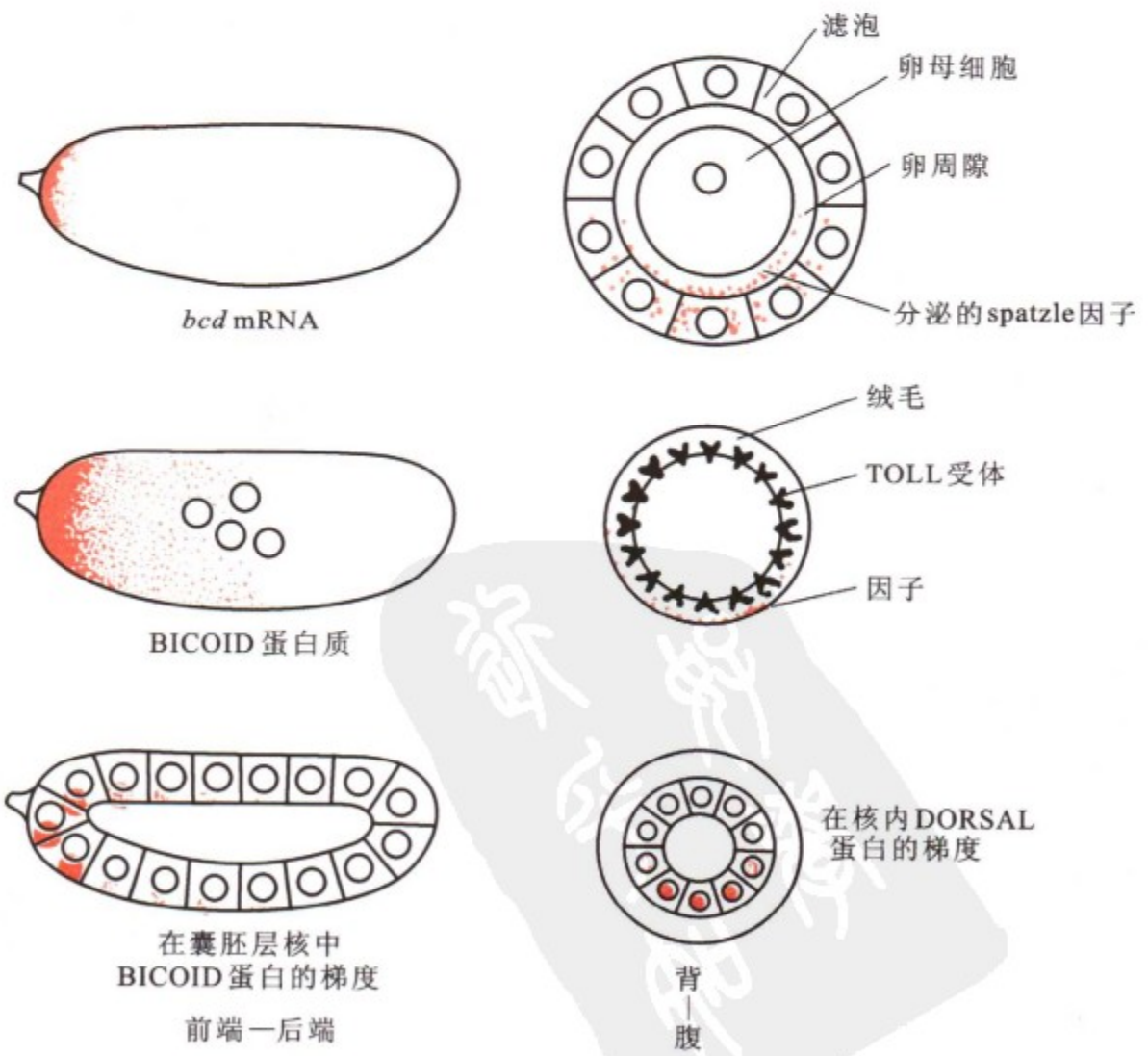



图 10–22 果蝇前后轴和背腹轴形态发生过程中的 BICOID（左）和 DORSAL（右）蛋白的分布

膜上或附近的母源效应基因 *spatzl* 的表达产物。因为只有在腹侧的受体能找到配基，所以，*spatzl* 前体定位于卵子的腹侧而不是背侧。被配基占据的受体能使 DORSAL 蛋白磷酸化从而引起 DORSAL 蛋白的重新分布，保证胚胎腹侧的细胞核中 DORSAL 蛋白浓度较高。当胚胎中 *DORSAL* 基因发生缺失时，胚胎不能产生腹侧结构，整个个体将呈现背侧外观。这个胚胎被称为背部化(dorsalized)胚胎。

### 10.2.4 影响果蝇体节发育的基因

果蝇躯体的分节是分步发生的。在合胞体囊胚层中，胚胎细胞开始转录和表达自身基因产物，其中有许多是转录调控因子。这些因子在果蝇身体内并不沿身体呈均匀分布，而是在空间上被限制在某些被称为表达区(expression zones)的区域内形成特殊的表达模式。目前已克隆了大约 25 个参与体节精细结构形成的基因。在囊胚层中，这些基因表达顺序如下：

1. 间隙基因

这类基因包括 hunchback(*hb*)、krüppel(*Kr*)和 Knirps(*kni*)等。这些基因的表达很有特点，最初在整个胚胎中都有很弱的表达，以后随着卵裂的进行而逐渐变转成一些不连续的表达区带(图 10-23)。

2. 成对规则基因

这些基因一般以两个体节为单位相互间隔一个副体节表达，其分布具有周期性。成对规则基因的功能是把间隙基因确定的区域进一步分化成体节。成对规则基因的表达是胚胎出现分节的最早标志之一，它们的表达模式较为独特，沿前-后轴形成一系列斑马纹状的条带分布，正好把胚胎分为预定的体节。

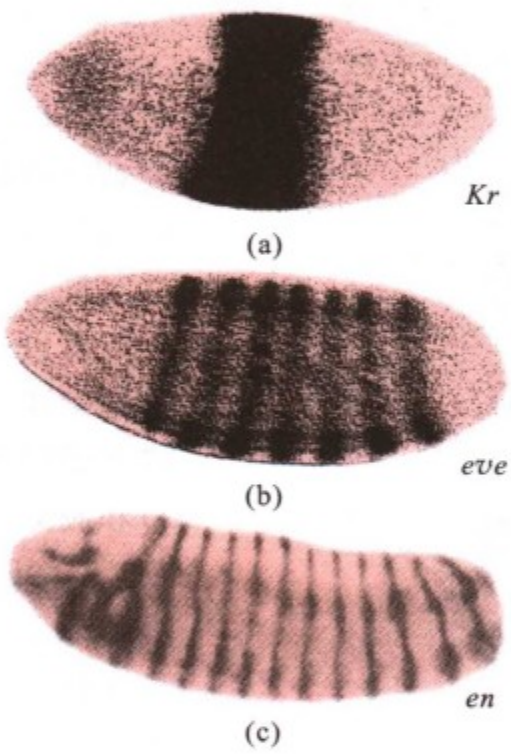
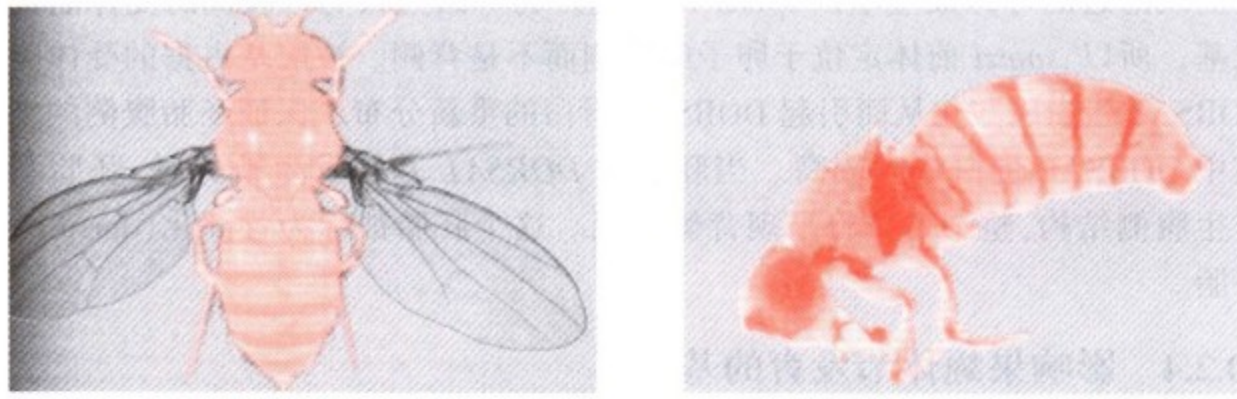



**图 10-23 三组参与胚胎发育早期体节分化基因的表达谱分析** (a)间隙基因(Krüppel,*Kr*)；(b)成对规划基因(*eve*)；(c)体节极化基因(engrailed,*en*)。

3. 体节极化基因

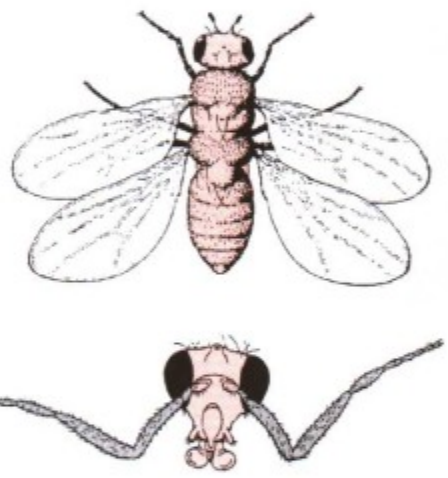
发育到细胞囊胚期时，体节极化基因(segment polarity genes)就把不同体节再进一步划分成更小的条纹(图 10-24)。当果蝇体节极化基因发生突变时，每个体节都会缺失一个特定的区域。engrailed(*en*)和 wingless(*wg*)基因是两个最重要的体节极化基因。*en* 基因在每一副体节最前端的一列细胞中表达，而 *wg* 基因的表达区域恰好位于 *en* 基因表达带之前。所以，这两个基因表达区域的界线正好是确立副体节的界线。*en* 和 *wg* 基因表达的起始受含有同源异型框的成对规则基因 *eve* 和 *ftz* 等编码转录调节因子的制约。*en* 基因在 *ftz* 和 *eve* 基因产物的浓度达到阈值后才能被激活，但 *wg* 基因的活化浓度显著低于这一阈值。



▌图 10-24 果蝇的体节特征(左)与用 ENGRAILED 抗体检测的该基因的表达(右)相吻合

### 10.2.5 果蝇的同源域基因

同源域基因最终决定躯体体节命运。躯体的部分最终是发育成为无翅的前胸还是有翅的中胸,是有平衡器的后胸还是腹部的体节,都是由同源域基因决定的。在果蝇中同源域基因统称 HOM 复合体(*HOM-C*)。大多数同源域基因位于第三号染色体上,排成两簇。一簇称为触角复合体(*Antp-C*);另一簇称为双胸复合体(*BX-C*)。同源域基因的存在是通过一系列非常引人注目的突变而得以证实的。果蝇的 *Antennapedia* 突变就是一个典型的例证。这种突变使果蝇的触角转变为足(图 10-25)。典型的同源域基因不参与基本躯干的建成,也不参与体节和肢芽的完善,它们只保证体节或肢芽的最终典型特征。若将调控形成头胸体节的 5 个 *Antp-C* 基因或参与胸腹体节形成的 3 个 *BX-C* 基因突变,可产生惊人的同源异型转换(homeotic transformation),使形态正确的结构长到了错误的地方。在触角足基因(*Antennapedia*)显性突变中,该基因在头以及胸部表达,使头结构部分转变成胸体节。因为该体节要承受的是腿而非触角,所以,头部出现了两条腿。



▌图 10-25 果蝇触角足复合物基因簇(Antennapedia complex)突变,引起同源异型转换 上,平衡器发育成翅膀(产生四翅果蝇);下,触角变成足。

## 10.3 高等植物花发育的基因调控

高等植物开花的控制长期以来是人们非常关心的事情,对这个问题的探索可追溯到 20 世纪初有关碳/氮比对开花迟早影响的学说,30 年代关于开花素的学说和春化的概念,70 年代所提出的关于植物由营养生长过渡到生殖生长前的感受状态理论。这个时期对开花的探索主要着重于植物的形态、生理学以及环境的影响等方面。也从遗传学角度利用植物本身的突变体或用人工创造的突变体对植物的开花提出了一些新的概念,并根据分生组织的形态特征,将植物从营养

生长到生殖生长的转变分为三个阶段，即营养分生组织阶段，花序分生组织发生阶段和花分生组织发生阶段。随着植物分子生物学技术的不断完善，尤其是有关克隆技术和诱变技术的出现，使人们从模式植物如拟南芥(*Arabidopsis thaliana*)或金鱼草(*Antirrhinum majus*)中分离克隆到许多控制植物开花的基因，为进一步认识植物开花的内在控制机制以及与外界环境因素的相互作用奠定了基础。

## 10.3.1 植物的花器官结构

种子植物的花由枝条变态产生，而花器官由叶片变态产生。一般认为，种子植物的成年花器官由花萼、花瓣、雄蕊、雌蕊(心皮)等四轮结构所组成(图 10-26)。成年花器官是从分生组织发育而来的，营养分生组织产生花序分生组织，花序分生组织产生花分生组织，花分生组织可被进一步分为早期花分生组织和晚期花分生组织，早期花分生组织有点类似于营养分生组织，具有无限生长的特点。花分生组织产生花器官原基，最后产生花器官。

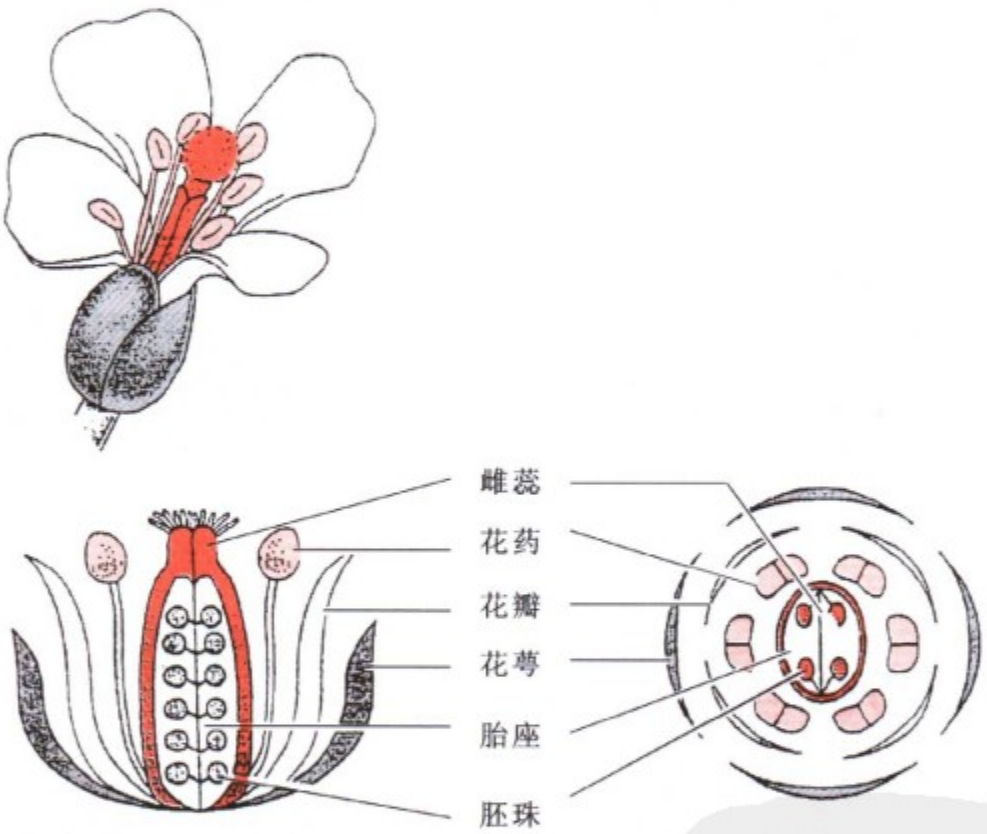



图 10-26 拟南芥花的构成

## 10.3.2 调控花器官发育的主要基因

花分生组织决定基因(flower meristem identity gene)促进从花序分生组织产生花分生组织并进一步分化产生花器官原基，但产生何种花器官则由同源域基因所控制。花分生组织决定基因可以在一定程度上激活同源域基因。

1. 同源域基因与花器官发育

金鱼草、拟南芥等植物的花都由四种类型的花器官组成，花器官排列成向心的圆环形，称作轮性(whorl)。野生型金鱼草和拟南芥的花由四个花萼组成最

外侧的第一轮、依次向内为四个花瓣组成第二轮，六个雄蕊组成第三轮，两个融合的心皮组成第四轮。

在研究金鱼草花形态突变体时发现，花器官的同源异型突变包含三大类型，类型Ⅰ为第一轮和第二轮器官受影响，产生心皮状的花萼和雄蕊状的花瓣。类型Ⅱ为第二轮和第三轮器官受影响，产生花萼状的花瓣和心皮状的雄蕊。类型Ⅲ为第三轮和第四轮器官受影响，产生花瓣状的雄蕊和花萼状的心皮，而且最内侧两轮器官的数量和轮数也发生了改变。类型Ⅳ中四轮结构全部受影响(图 10–27)。进一步研究这些突变体的基因型发现，每一种类型的突变体中都

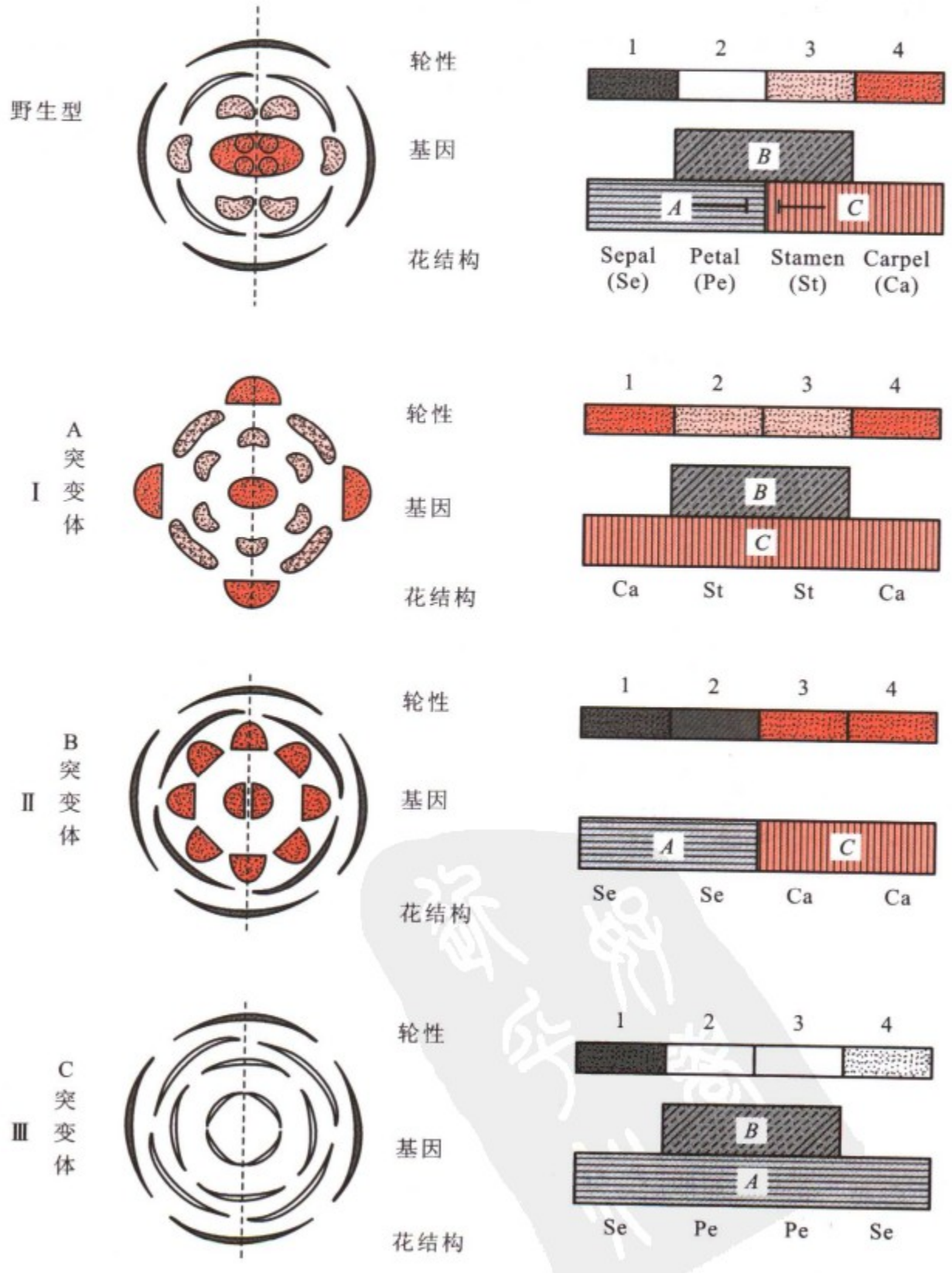



图 10–27 由 3 组同源异型基因决定四轮花器官特征的"*ABC*"模型

是因为发生了同源域基因的突变而使相邻两轮花器官受到影响,证明植物的器官发生与动物的器官发生一样,都受同源域基因控制。

2. 同源域基因的分离

科学家最早在 1990 年利用转座子突变的方法从金鱼草中分离并克隆出同源域基因 *DEFA*(Deficiens)。研究发现,该基因编码的蛋白质与两个已知的转录因子,哺乳动物中的 SRF(serum response factor)和酵母中的 MCMI(minichromosome maintenance gene)的一个保守区域有很高的同源性,这一区域参与二聚化并与 DNA 结合,在推测的与 DNA 的结合区中,单个的氨基酸突变均会导致该转录因子与 DNA 的结合能力下降,说明 *DEFA* 可能编码一个具有调控功能的 DNA 结合蛋白。

此后,又克隆了同源域基因 *AG*(agamous),它编码的蛋白质与上述转录因子具有高度的序列相似性。用 *DEFA* 的这一保守区域作探针,从金鱼草的 cDNA 库中筛选出了 8 个独立的基因,在蛋白质水平上与 *DEFA* 的 DNA 结合区有很高的相似性(65%~90%)。表 10-4 列出了已克隆的多个涉及花发育的同源域基因。上述结果表明,在金鱼草和拟南芥中可能存在着一个参与花器官发育和分化的基因家族,该家族的成员可能参与不同的分化过程。由于这些新的家族(*MCM1*、*AG*、*DEFA* 和 *SRF*)均含有一个保守的区域,根据这几个家族基因名称的第一个字母,将该保守区命名为 MADS-box。

表 10-4 已经克隆的植物花器官特征决定基因及其可能的功能分析

| 基因(拟南芥/金鱼草) | 突变体表型 | 基因功能 |
|---|---|---|
| APETALA1/SQUAMOSA | 花萼变为叶片 | 转录因子 |
| APETALA2/Unkown | 花萼变为叶片/心皮,花瓣变为心皮 | 转录因子 |
| APETALA3/DEFICIENS | 花瓣变为雄蕊,雄蕊变为心皮 | 转录因子 |
| PISTILLATA/GLOBOSA | 类似 AP3/DEF | 转录因子 |
| AGAMOUS/PLENA | 重复发生一轮花萼及两轮花瓣 | 转录因子 |
| LFY/FLORICAULA | 产生更多的花序,绿色花,由类似花萼与心皮的花器官构成 | 转录因子 |

## 10.3.3 花器官发育的“ABC”模型

1. “ABC”模型的提出

在对拟南芥和金鱼草突变体及花器官特征决定基因功能的研究中,E. Myerowitz 提出了控制花形态发生的“ABC”模型。图 10-27 是以拟南芥突变体为代表的“ABC”模型图示。根据这个模型,正常花的四轮结构的形成是由三组基因共同作用而完成的。每一轮花器官特征的决定分别依赖于 *A*、*B*、*C* 三组基因中的一组或两组基因的正常表达。如其中任何一组或更多的基因发生突变而丧失功能,则花的形态将出现异常。*A* 基因在第一、二轮花器官中表达,*B* 基因在第二、三轮花器官中表达,*C* 基因在第三、四轮花器官中表达。*A* 基因本身决定

萼片，*A* 和 *B* 基因共同决定花瓣，*B* 与 *C* 基因共同决定雌蕊，*C* 基因决定心皮。此外，*A* 基因与 *C* 基因相互拮抗。当 *A* 基因突变后（如 *AP1*、*AP2*），导致第一轮花器官中的花萼突变为心皮，第二轮花器官中的花瓣突变为雄蕊。*B* 基因突变后（如 *AP3*、*PI* 突变体），花萼替代了第二轮花器官中的花瓣，第三轮花器官中的雄蕊变为心皮。若 *C* 基因失活（如 *AG* 突变体），则第三轮花器官中的雄蕊转变为花瓣，第四轮花器官中的心皮也转变为花瓣。*B* 基因在第四轮花器官中的表达受到 *SUPERMAN* 基因的抑制。*SUP* 突变体的表型为第四轮心皮数量大大减少，而雌蕊数量大大增加。如前三轮花器官中的基因表达正常，只有第四轮花器官中有 *B* 基因异位表达，导致该轮花器官发育为雄蕊（图 10-27）。大量单突变、双突变、三突变体以及基因过量表达的研究结果均在不同程度上支持这个模型。“ABC”模型的提出是近年植物发育生物学研究中的一个非常重要的突破，可以解释多个基因在器官特征性发育中的作用。

2. 对“ABC”模型的质疑

“ABC”模型是基于科学家对所分离的控制花器官决定基因作用方式的一种简单化的概念性解释，随着人类在分子水平上对同源域基因相互作用的新认识，同源域基因所调控的“ABC”模型也受到了许多质疑。首先，在“ABC”模型中不同类型器官的划分是很严格的，但许多开花突变体中，不同类型器官之间常常会出现嵌合体；其次，*AG* 突变体在理论上应使第三四轮组织发育成为相同数量的花萼和花瓣，事实上往往出现增殖现象，产生大量花萼花瓣，暗示 *AG* 可能不是一个单独的基因，而是有多个同源基因，这些基因不但在雄蕊和心皮发生过程中起作用，也在其他生命过程中起作用；第三，*AP2* 基因的表达不具备器官特异性，所有花器官中都存在该基因的表达产物。最后，*SUPERMAN* 基因能抑制 *AP3* 和 *PI* 基因表达，从而抑制花器官形成（图 10-28）。最近，有人提出了控制花形态发生的“*ABCE*”模型，即除在控制花形态中的 *ABC* 基因之外，还有一类在功能上重复的基因——*E*。这类基因是由多个相似基因编码相关蛋白与 *ABC* 基因共同控制特异花器官如花萼，花瓣，雄蕊和心皮的形成，完全缺失 *E* 基因则形成叶状器官，*ABC* 基因失去形成花器官的功能，不能产生花器官，如*sep1sep2sep3* 三突变体就只有花萼（图 10-29）。

### 10.3.4 参与诱导开花信号产生、传递和感知的基因

已从拟南芥中克隆了两个涉及开花时间的基因，CONSTAINS（*CO*）和 LUMINIDEPENDENS（*LD*）。研究表明，这两个基因都在营养生长期表达。CO 可能是一个转录因子，一般情况下促进长日照生长的植物开花。*CO* 基因的异位表达会迅速诱导开花基因的表达，使植物开花时间大大提前（图 10-30）。

最终促进开花信号来源于花器官分生组织决定基因，LEAFY（*LFY*），目前尚不清楚该基因表达的产物是否直接与促进开花的信号发生相互作用。拟南芥 *LFY* 突变体比野生型产生更多的花序分枝，其花呈绿色，由类似花萼和心皮样的器官构成，异位表达 *LFY* 导致转基因植物提前开花并将茎尖转变成花，证明 *LFY* 基因不仅决定花分生组织的特性，而且影响开花时间（图 10-31）。

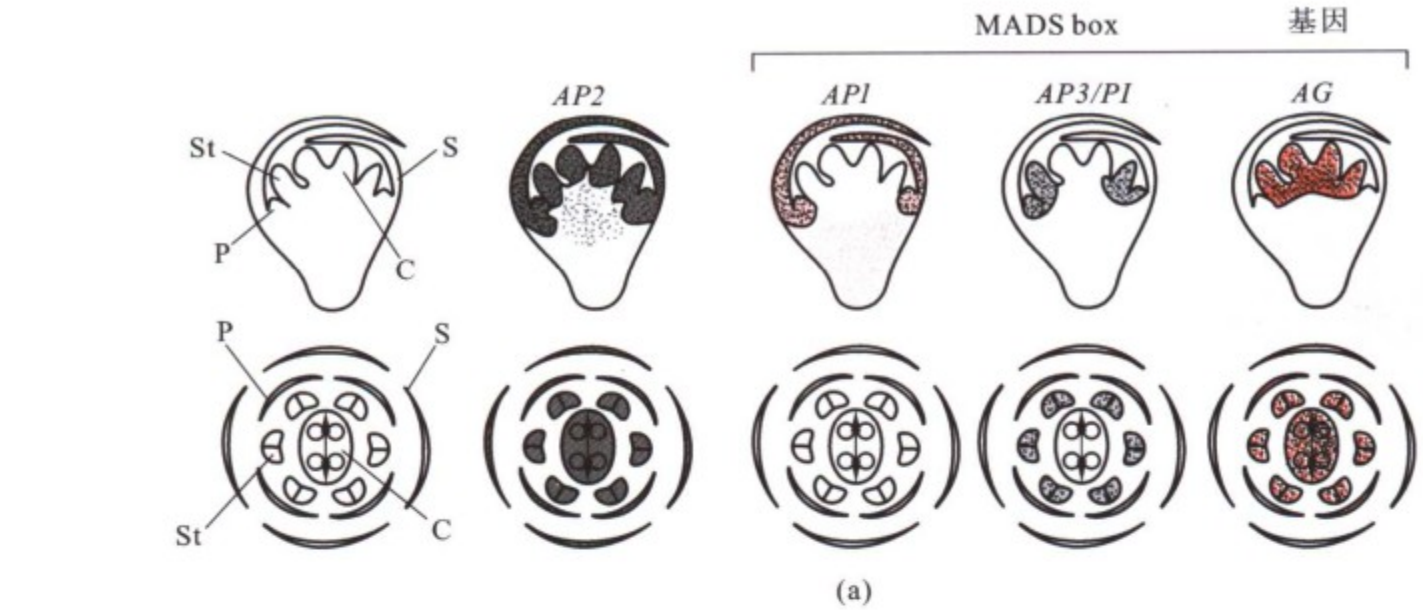



**图 10-28 经修改后的"*ABC*"模型** (a)*AP2* 基因和花同源异型 MADS 盒基因与植物花发育的关系分析。AP2在四轮花器官中都表达,MADS box 基因只限定在一对花器官中表达。S,花萼;P,花瓣;St,雄蕊;C,心皮(雌蕊)。(b)通过对相关基因表达区域分析修正"*ABC*" 模型。*X* 基因尚未被克隆。

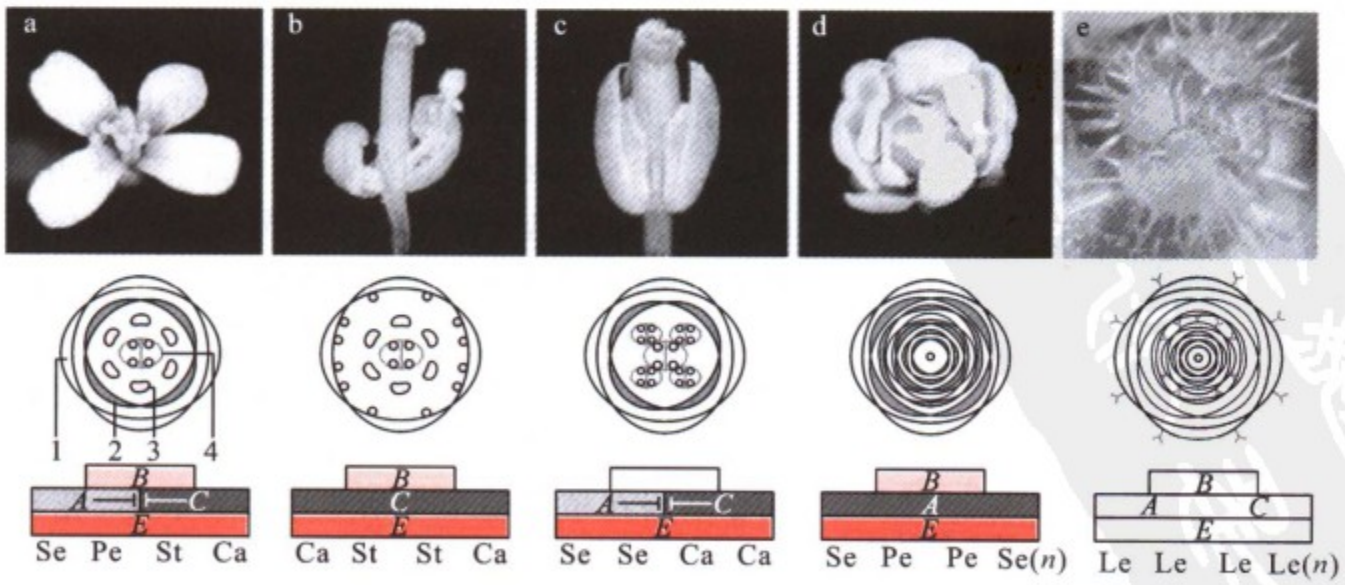



**图 10-29 花形态建成的 *ABCE* 模型** a. 野生型;b. *A* 基因突变;c. *B* 基因突变;d. *C* 基因突变;e. *E* 基因突变。

(a) (b)

**图 10–30 *CO* 表达诱导拟南芥提前开花** (a)野生型拟南芥;(b)35 S 启动子驱动 *CO* 基因表达,诱导拟南芥提前开花。

*TERMINA1 FLOWER1*(*TFL1*)是影响拟南芥分生组织特性的另一个非常重要的基因。*TFL* 突变体开花提前,初级花序分枝转变成末端花,缺少侧枝,花的数量也大大减少(图 10–32)。*TFL1* 在分生组织中表达,但不在幼花原基中表达。*TFL* 突变体(由无限枝转变成有限的枝)和 *LFY* 突变体(由有限的枝转变成无限的枝)的表型恰好相反,说明两个基因可能是相互拮抗的。

另外三个影响花分生组织决定基因是 *AP1*、*AP2*和 *CAL*。*AP1* 和 *AP2* 突变强化了 *LFY* 突变体的表型,虽然 *CAL* 突变体没有可见的表型,但 *CAL* 和 *AP1* 双突变体转变成更明显的无限枝条,产生了类似与菜花花型样的花序,说明该基因对花分生组织的发生有重要影响。*AP1* 和 *CAL* 基因序列和表达模式非常相似,说明它们可能在某种程度上重叠或互补。

*LFY*、*AP1* 和 *AP2* 表达模式与它们在早期花分生组织中的作用相吻合。在花器官原基发育之前,*LFY* 和 *AP1* 在营养生长过程中表达量很低,在 *GA* 或光周期处理诱导开花过程中,*LFY* 表达量会迅速提高。*CO* 的异位表达也会促进 *LFY* 的大量表达,说明 *CO* 的功能之一是促进 *LFY* 表达。虽然在由营养生长转变到花发育时 *LFY* 基因已经表达,但 *AP1* 在花原基中没有表达,直到花诱导开始 1~2 d 后才开始表达。*AP2* 基因在植物整个生活史中持续表达,花器官分生组织确定期该基因表达量显著上升(图 10–33)。

*LFY*、*AP1* 和 *AP2* 基因具有不同的功能,但

(a)

(b)

**图 10–31 花分生组织决定基因 *LEAFY*(*LFY*)将具有无限生长特征的拟南芥茎尖组织转变成花** (a) 在 *LFY* 突变体中正常情况下发育成花的枝转变成无限枝条。图中所示的所有次生枝条生长的位置在野生型拟南芥中均发育成花序;(b) 过量表达 *LFY* 的转基因拟南芥中,次生枝转变成花,初级枝在早期发育成终端花或形成一簇花。

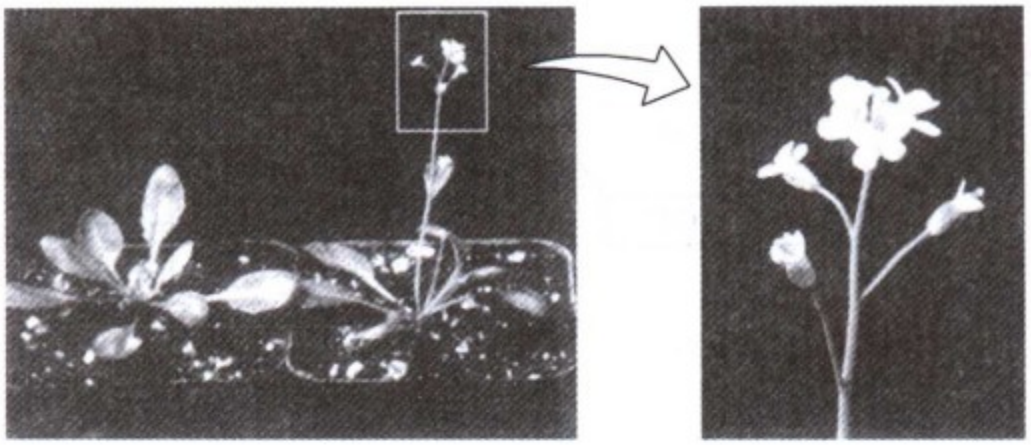



**图 10-32 *TFL1* 对花发育的调控** (a)*TFL* 突变体的花序表型。左边为野生型,右边为 *TFL* 突变体,很快就发育成一个终端花或一簇终端花;(b)*TFL* 突变体终端花的放大图。

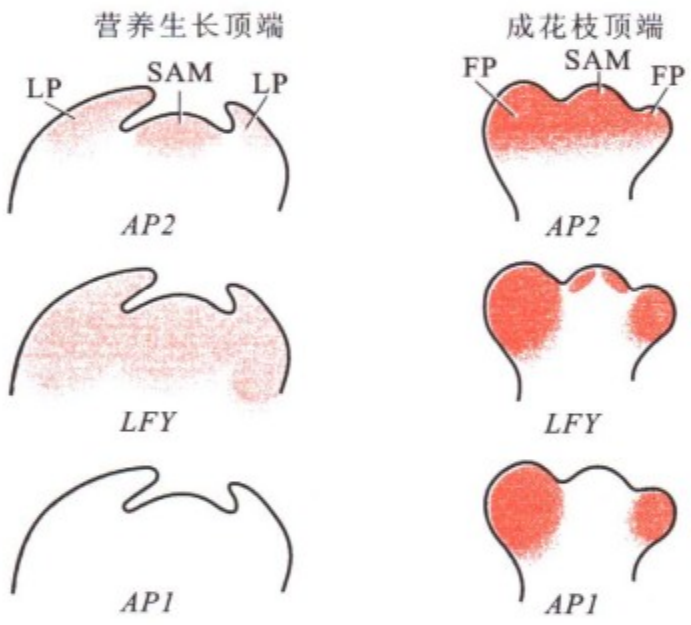



**图 10-33 花分生组织决定基因在营养生长和成花枝顶端的表达模式分析** LP,叶原基;SAM,顶端分生组织;FP,花原基。

它们在花发育不同阶段有重叠,推测可能有增强基因活性的作用,因为这些基因都编码了转录调控因子。序列分析表明,*AP1*、*AP2* 类转录因子含有 MADS 盒,*LFY* 编码的蛋白质富含脯氨酸和酸性区,它可能代表着一类新的转录因子。图 10-34 是各条开花通路之间的相互关系,许多科学家正在共同努力,争取尽快全面揭示植物开花的调控规律,进一步造福人类。

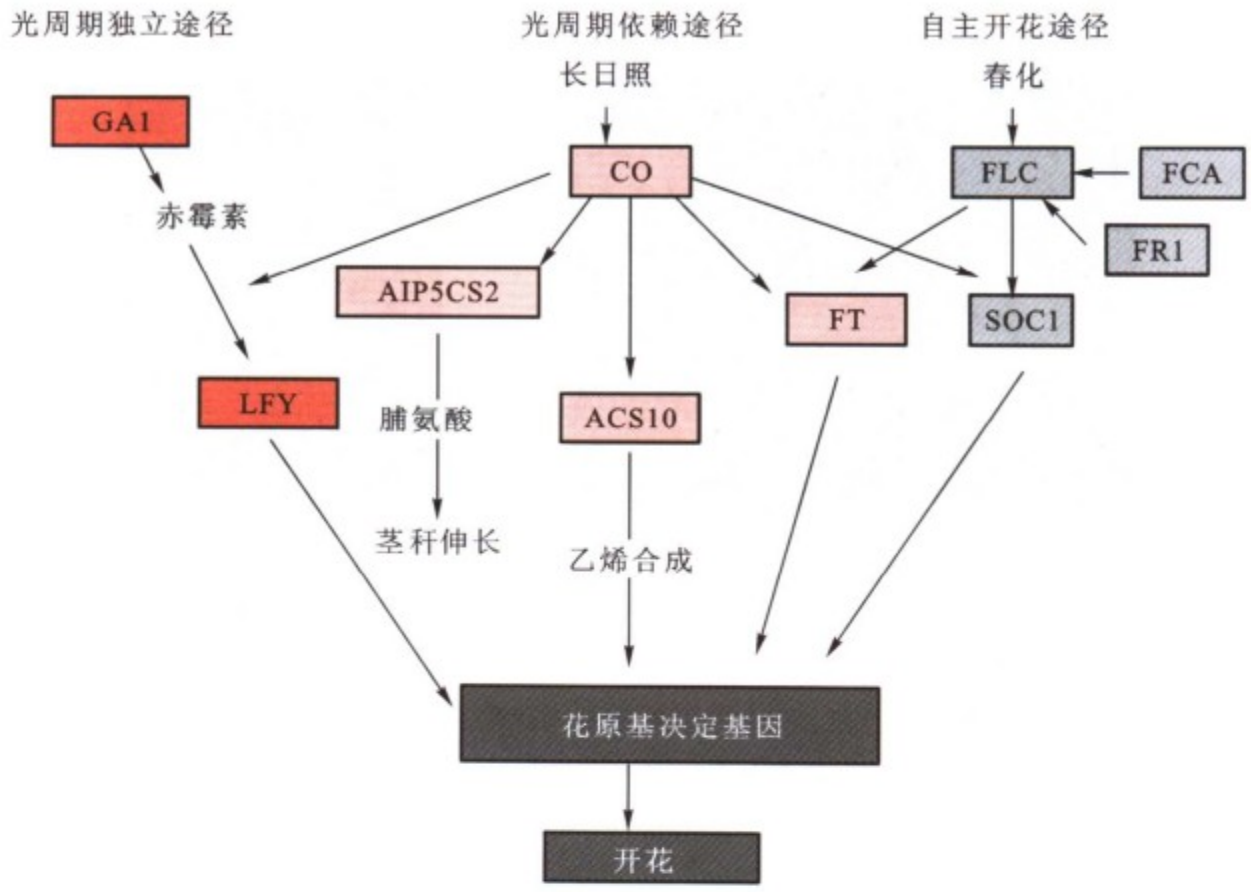



图 10-34 拟南芥中各条开花通路之间的相互关系分析

## 思考题

1. 为何机体需要体液免疫和细胞免疫两种免疫方式?缺失其中一种的后果是什么?
2. 假如你猜测某个基因涉及果蝇的眼睛发育(已知该基因的序列),请试设计实验证实你的设想。
3. 假如某个基因突变使果蝇长出两个尾巴，但你不知道该表型涉及一个基因还是多个基因,你将如何设计实验解决这一问题? 请提出实验方案找到涉及这个性状的突变基因。
4. 假如你得到一个拟南芥突变体,表型是花发育不正常,但是发生突变的基因不属于所有已知的参与花形态建成的基因，你如何研究这个基因在拟南芥花发育中的功能? 你觉得这个问题有可能吗(即确实存在这样的基因吗)?
5. 现在已经证明植物开花的时间与光照有关，试问这种关联对于植物的意义在哪儿?
6. 假如把拟南芥中所有涉及花发育的基因全部敲除，将所有涉及玫瑰花发育的基因都置于相应拟南芥基因启动子的调控之下并转到拟南芥中，这些转基因拟南芥会开出玫瑰花吗? 为什么?

## 主要参考文献

1. Krizk B A, Fletcher J C. Molecular mechanisms of floral development: an arm-

chair guide. Nature Review Genetics. 2005,6:688-698.

2. Bäurle I,Dean C. The timing of developmental transitions in plants. Cell,2006, 125:655-664.

3. Bennasser Y,Le S Y,Benkirane M. Evidence that HIV-1 encodes an siRNA and a suppressor of RNA silencing. Immunity,2005,22:607-619.

4. Thompson B J,Cohen S M. The Hippo pathway regulates the bantam microRNA to control cell proliferation and apoptosis in Drosophila. Cell,2006,126:767-774.

5. Mara M A,Jones S J,Astell C R. *et al*. The genome sequence of the SARS-associated coronavirus. Science,2003,300:1399-1404.

# 11 基因组与比较基因组学


- 11.1 高通量 DNA 序列分析技术
- 11.2 人类基因组计划
- 11.3 其他基因组
- 11.4 比较基因组学及相关研究


20 世纪 40 年代第一颗原子弹爆炸,60 年代人类首次登上月球和 90 年代提出并已基本完成的人类基因组计划(human genome project,HGP)是这 100 年科技发展史上的三大创举。基因组是指生物有机体的单倍体细胞中的所有 DNA,包括细胞核内的 DNA 和各种细胞器中的 DNA。DNA 双螺旋结构的发现者之一 J.D.Watson 认为,与人类登月计划相比,HGP 的资金投入少,但它对人类生活的影响却可能更深远。因为随着这个计划的完成,DNA 分子中储藏的有关人类生存和繁衍的全部遗传信息将被破译,它将不仅帮助我们理解人类如何作为健康人发挥正常生理功能,还将最终揭开基因在癌症、早老性痴呆症、精神分裂症等严重危害人类健康的疾病中的作用。事实上,对人类自身更深入的了解是人类活动最重要的组成部分,因为任何自然科学研究,都不会比人类尽快找出解决自身所面临的人口膨胀、粮食短缺、环境污染、疾病危害、能源资源匮乏、生态平衡破坏、生物物种消亡等一系列难题更为重要、更为迫切。

对人类以及其他生物体全基因组的测序是人类科学发展史上的一项浩大工程。人类基因组计划以及由此派生出来的数量庞大的基因组计划显然得益于 20 世纪下半页获得蓬勃发展的分子生物学研究和以“微软”技术为龙头的计算机软件行业,前者通过发明 DNA 序列分析仪器和迅速的更新换代为基因组计划提供了强大的技术支撑,而后者使该计划所产生的海量数据在实验室得到消化和处理。从 1975 年花了九牛二虎之力测定全长 5 386 bp 的噬菌体 ΦX174 基因组到 2001 年完成总长 $3\times10^9$ bp 的人类基因组序列测定,短短 20 多年间 DNA 序列分析技术发生了质的飞

跃。如果没有高通量 DNA 序列分析技术的突破,人类基因组计划就几乎不可能被完成。有人做过这样的统计,如果一个熟练的DNA 序列分析人员采取早期方法每天测定 1 000 bp 计算,人类基因组(约 $3\times10^9$ bp)的全序列分析,至少也得需要 100 名这样的工人花上 100 年的时间才能完成。而现在的高通量测序仪可达到月测序 1 Mb~6 Mb! 随着计算机分析处理能力的提高以及科学家对序列分析流程的不断改进和优化,必将有越来越多物种的基因组 DNA 被测定和解析。

## 11.1 高通量 DNA 序列分析技术

人类基因组计划的成功很大程度上得益于早期在有效减少 DNA 测序成本的方法学上的投资。通过改良测序方法,不断提升其自动化程度,DNA 测序的成本降低了 100 倍,从 20 世纪 70~80 年代每个碱基 10 美元降低到目前 1 美元 10 个碱基! 可以说,高通量测序仪器的引入和大规模商业化测序中心的建立推动了新的超低成本 DNA 测序技术的发展。在可以预见的将来,很可能有更低成本的 DNA 测序技术走向市场。

### 11.1.1 DNA 序列测定的基本原理

无论是对于重组质粒、单个基因还是整个基因组,分析 DNA 结构的最基本方法就是测定出这些 DNA 分子的一级结构——DNA 序列。DNA 自动测序仪的出现使得 DNA 测序过程变得快捷而稳定, 计算机处理能力的迅速提高则使得大量DNA 小片段很容易被拼接成较大的片段甚至整条染色体。高效快捷的 DNA 测序方法是 20 世纪 70 年代中期发展起来的。当时有两种不同的测序方法,一种是 Sanger 的双脱氧链终止法,一种是 Maxam-Gilbert 的化学修饰法。在基因组测序的过程中双脱氧链终止法逐渐占据了显著的优势,因为该方法更适合于序列分析的自动化。

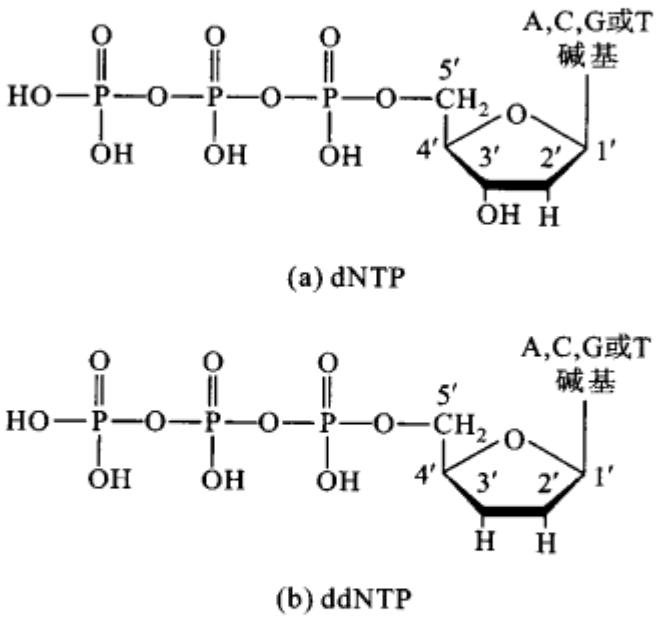



**图 11-1 脱氧核苷三磷酸(dNTP)和双脱氧核苷三磷酸(ddNTP)的分子结构式**

剑桥大学的 F.Sanger 等人于 1977 年发明了利用 DNA 聚合酶的双脱氧链终止原理测定核苷酸序列的方法。由于这种方法要求使用一种单链 DNA 模板和一种适当的 DNA 引物,因此也称为引物合成法或酶催化引物合成法。它利用了 DNA 聚合酶所具有的两种酶催化反应的特性:①该酶能以单链 DNA 为模板, 合成出准确的 DNA 互补链序列; ②如果以 2′,3′-双脱氧核苷三磷酸为底物,掺入到新合成的寡核苷酸链的 3′端后,DNA 链的延伸被终止(图 11-1)。在 DNA 测序反应中, 加入模板

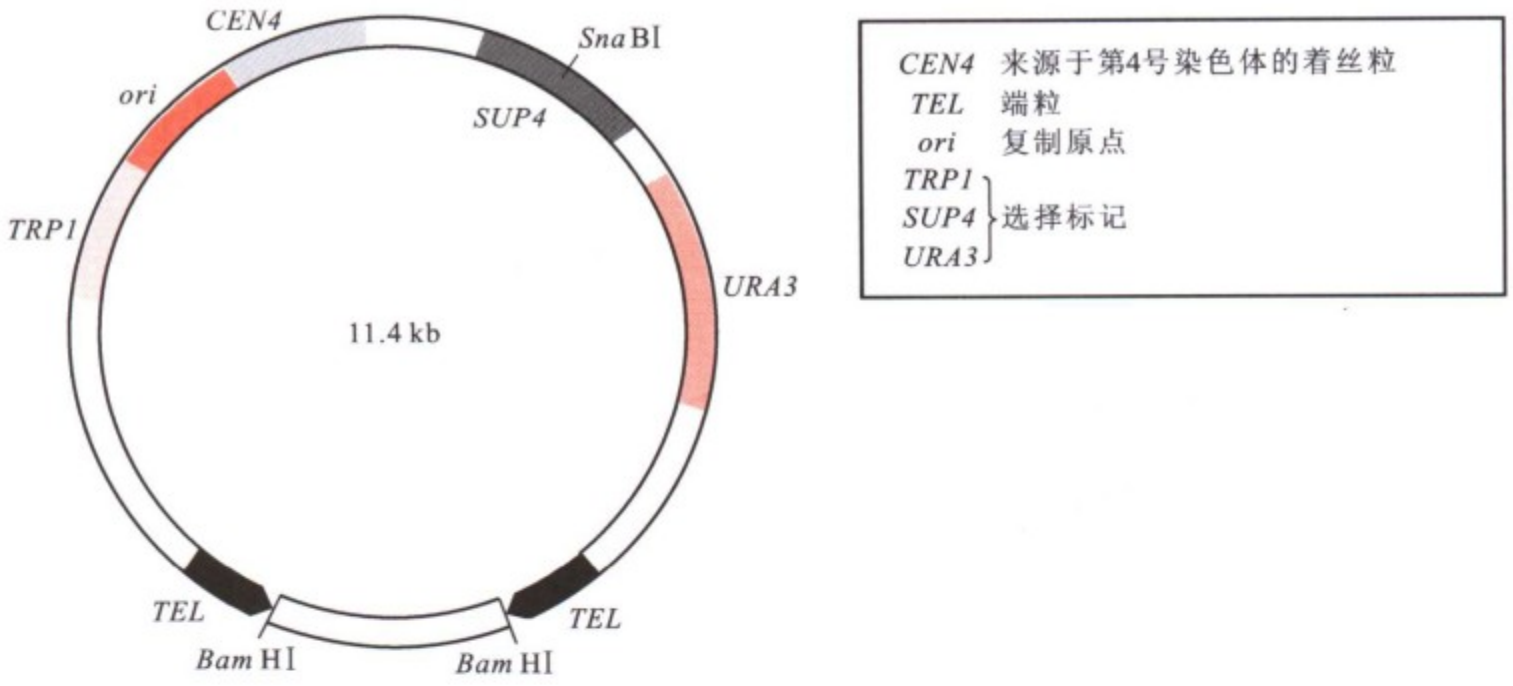
DNA、特异性引物、DNA 聚合酶、dATP、dTTP、dGTP、dCTP 和一种 ddNTP，当这种 2′,3′-双脱氧 ddTTP 掺入到寡核苷酸链的 3′端，取代了常规的脱氧核苷酸(dNTP)之后，由于 ddTTP 没有3′-OH 基团，寡核苷酸链不再继续延长，在本该由 dTTP 掺入的位置上发生了特异性的链终止效应。如果在同一个反应中，加入带 $^{32}P$ 放射性标记的 dNTP，经过适当的温育之后，将会产生出不同长度的 DNA 片段混合物。它们都具有同样的 5′端，并在3′端的 ddTTP 处终止。将这种混合物加到变性凝胶上进行电泳分离，就可以获得一系列全部以 3′端 ddTTP 为终止残基的 DNA 片段的电泳谱带模式。分别加入 ddATP、ddGTP 和 ddCTP，在不同试管中温育后，点样于同一变性凝胶上做电泳分离，再通过放射自显影的方法检测单链 DNA 片段的放射性带，就可以直接读出 DNA 的核苷酸顺序。在 DNA 合成反应中，一般用去掉 5′→3′核酸外切酶活性的 DNA 聚合酶Ⅰ的 Klenow 大片段来催化合成单链 DNA 模板序列的互补链。

### 11.1.2 基因组 DNA 大片段文库的构建

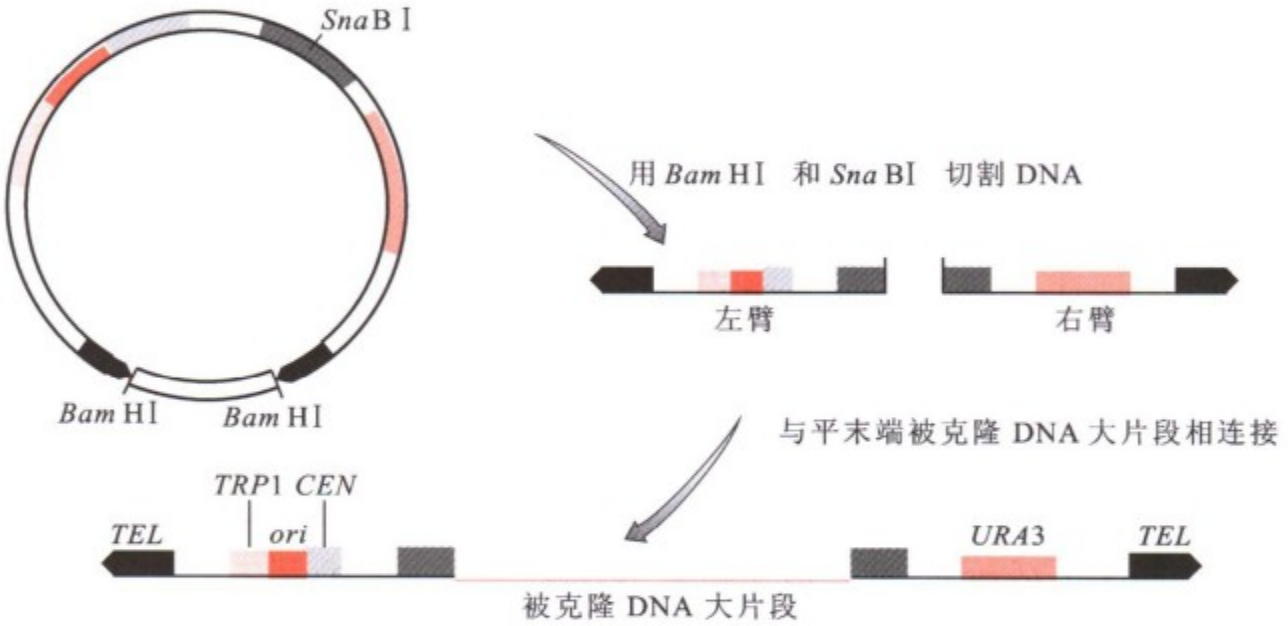
除了 DNA 序列测定技术进步之外，20 世纪 80 年代后期发展起来的酵母人工染色体技术(yeast artificial chromosome，YAC)为创制基因组物理图以及最终的基因组序列测定提供了一个极为方便的平台。1983 年，美国 Dana-Farber 癌症研究所和哈佛大学医学院的科学家首次报道了构建 YAC 库的过程。1987 年，科学家发现，仅仅带有 ARS 序列的载体虽然能够被复制，但极易在有丝分裂时丢失。即使在选择培养基上，也只有 5%~20%的子代细胞带有 ARS 类载体。若在 YAC 载体上加入(CEN)序列，就能显著提高有丝分裂时的稳定性，保证 90%以上的子代细胞带有该载体。CEN 还能显著降低拷贝数，使每个细胞所带的载体数从 20~50 个降为 1~2 个。

YAC 是迄今容量最大的克隆载体，插入片段平均长度为 200~1 000 kb，最大的可以达到 2 Mb。YAC(图 11-2)含有三种必需成分：着丝粒、端粒和复制起点。着丝粒(CEN)位于染色体中央，呈纽扣状结构，在有丝分裂时结合微管并调控染色体的运动，也是姐妹染色单体配对时的最后位点，接收细胞信号而使姐妹染色体分开。端粒(TEL)的主要功能是防止染色体融合、降解并确保其完整复制。但是，YAC 载体也存在不少缺点，特别是其嵌合体比例较高，一个 YAC 克隆中可能含有两个或多个本来不相连的独立的 DNA 片段。部分克隆子不稳定，在继代培养过程中可能会发生缺失或重排。由于 YAC 与酵母染色体具有相似的结构，实验操作中很难与酵母染色体区分开，操作时也容易发生染色体机械切割和断裂。

为了克服上述缺点，科学家用细菌的 F 质粒及其调控基因构建了细菌染色体克隆载体，称为 BAC(bacterial artificial chromosome)，其克隆能力在 125~150 kb 左右。该质粒主要包括 *oriS*，*repE*(控制 F 质粒复制)和 *parA*、*parB*(控制拷贝数)等成分(图 11-3)。以 BAC 为基础的克隆载体形成嵌合体的频率较低，转化效率高，而且以环状结构存在于细菌体内，易于分辨和分离纯化，已被科学界广泛接受。




(a) pYAC3




(b) pYAC3 的克隆策略

图 11-2 酵母人工染色体克隆载体及其克隆策略示意图

### 11.1.3 鸟枪法序列测定技术及其改良

受 DNA 序列分析技术的限制，一次测序反应的长度不能超过 1 000 bp，目前尚不能直接用 BAC 等大片段作为序列分析的模板，而往往采用所谓的全基因组鸟枪法测序技术，随机挑选带有基因组 DNA 的质粒做测序反应，然后用计算机程序进行序列拼接(图 11-4)。

全基因组鸟枪法测序的基本原理是：对某基因组文库全部克隆片段进行末端序列测定中未测到的碱基数，即缺口(gap)，与已测定的总碱基数相关。随着已测定碱基数的增加，缺口的总碱基数目会按照泊松公式的一个推论($P=e^{-m}$)迅速减小。其中 $P$ 为基因组中某个碱基未被测定的概率，$m$ 为所测定

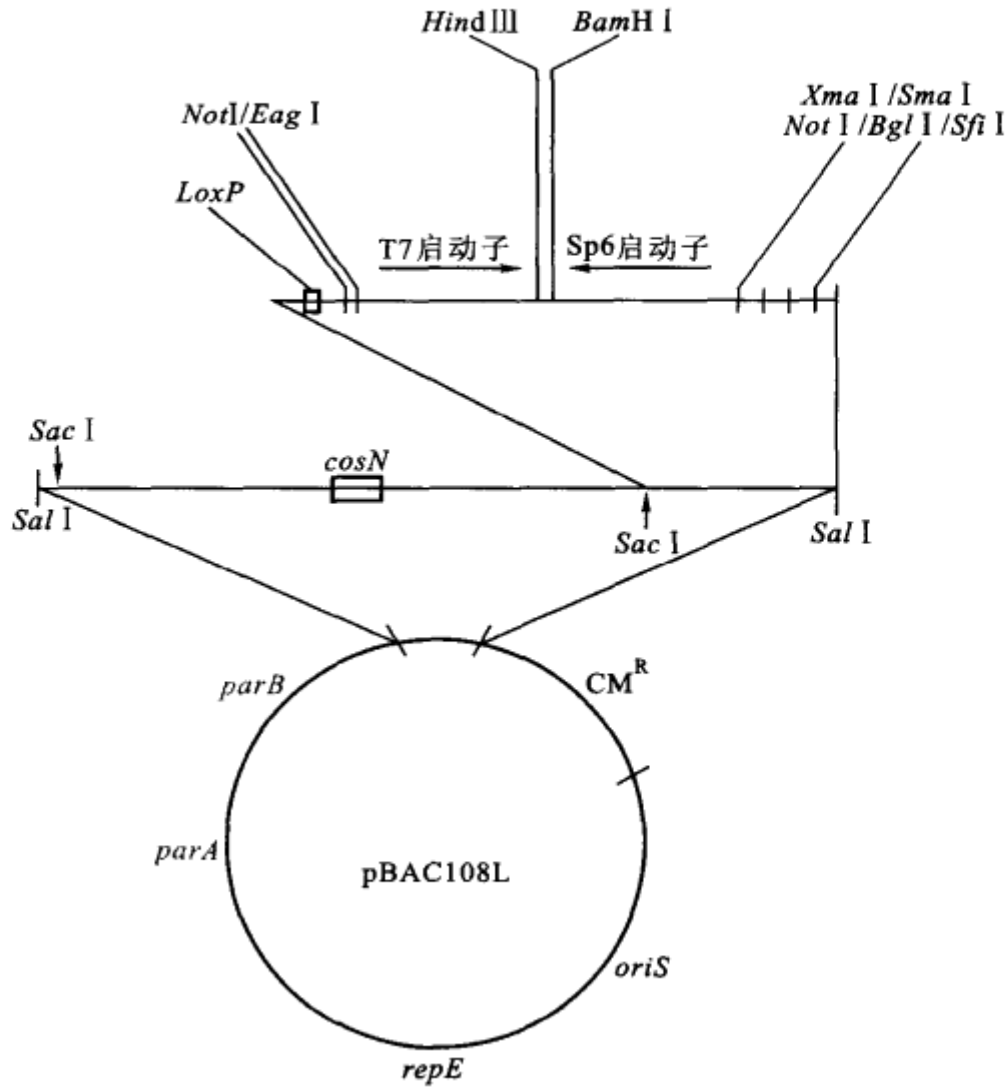



**图 11-3 细菌人工染色体的构建及其克隆策略** pBAC108L 来自细菌的一个小型 F 质粒，其中 *oriS* 和 *repE* 控制了质粒的复制起始，*parB* 和 *parA* 控制了拷贝数。

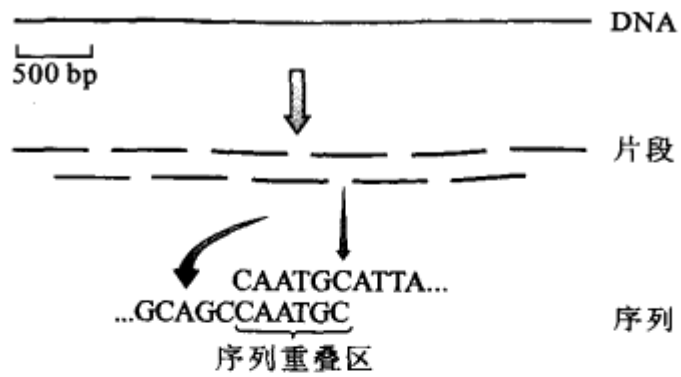



**图 11-4 基因组 DNA 的鸟枪法测序原理示意图**

的碱基数与基因组大小相比的倍数，$m$ 越大 $P$ 值越小。当 $m$ 值达到 5（即随机测定的碱基数达到基因组 5 倍时），基因组中未测定的碱基数为基因组总碱基数的 0.67%（$e^{-5}$= 0.006 7）。对流感嗜血杆菌这样大的基因组（1.83 Mb），此时可能仍有 128 个平均长度为 100 bp 的缺口。第一，要建立高度随机、插入片段大小为 1~2 kb 的基因组文库。克隆数要超过一定值，并保证经末端测序的克隆片段的碱基总数达到基因组 5 倍以上。

第二，高效、大规模的末端测序。对文库中每一个克隆，进行两端测序，TIGR 在完成流感嗜血杆菌的基因组时，使用了 14 台测序仪，用三个月时间完成了必需的 28 463 个测序反应，测序总长度达 6 倍基因组。

第三，序列集合。TIGR 发展了新的软件，修改了序列集合规则以最大限度

地排除错误的连锁匹配。

第四，填补缺口。有两种待填补的缺口，一是没有相应模板 DNA 的物理缺口，二是有模板 DNA 但未测序的序列缺口。他们建立了插入片段为 15~20 kb 的 λ 文库以备缺口填补。

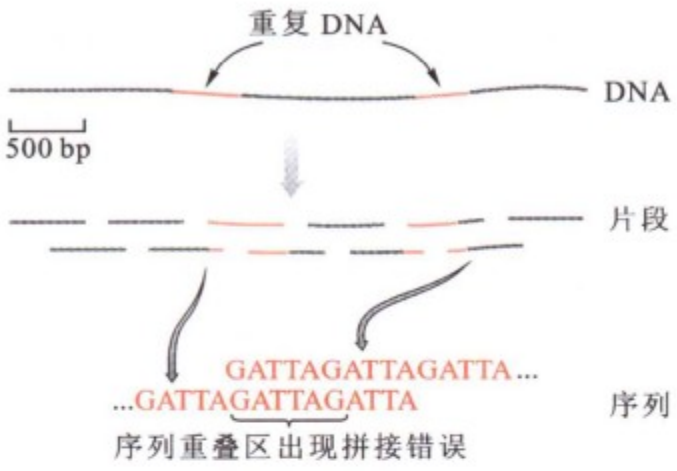



图 11-5 鸟枪法测序技术不能鉴别高等真核生物基因组中的重复序列

鸟枪法测序的主要缺点是，随着所测基因组总量增大，所需测序的片段大量增加，各个片段重叠成一个连续体的概率是 $2n^2-2n$。其次，高等真核生物（如人类）基因组中有大量重复序列，导致判断失误（图 11-5）。

为了克服上述问题，现在大多采用大片段克隆法（clone contig）或靶标鸟枪法（direted shotgun）进行体积较大的真核基因组序列分析。大片段克隆法首先用稀有限制性核酸内切酶把待测基因组降解为几十万个碱基对的片段，再分别测序，而靶标鸟枪法则根据染色体上已知基因或遗传标签的位置来确定部分 DNA 片段的排列顺序，再逐步确定各片段在染色体上的相对位置（图 11-6）。

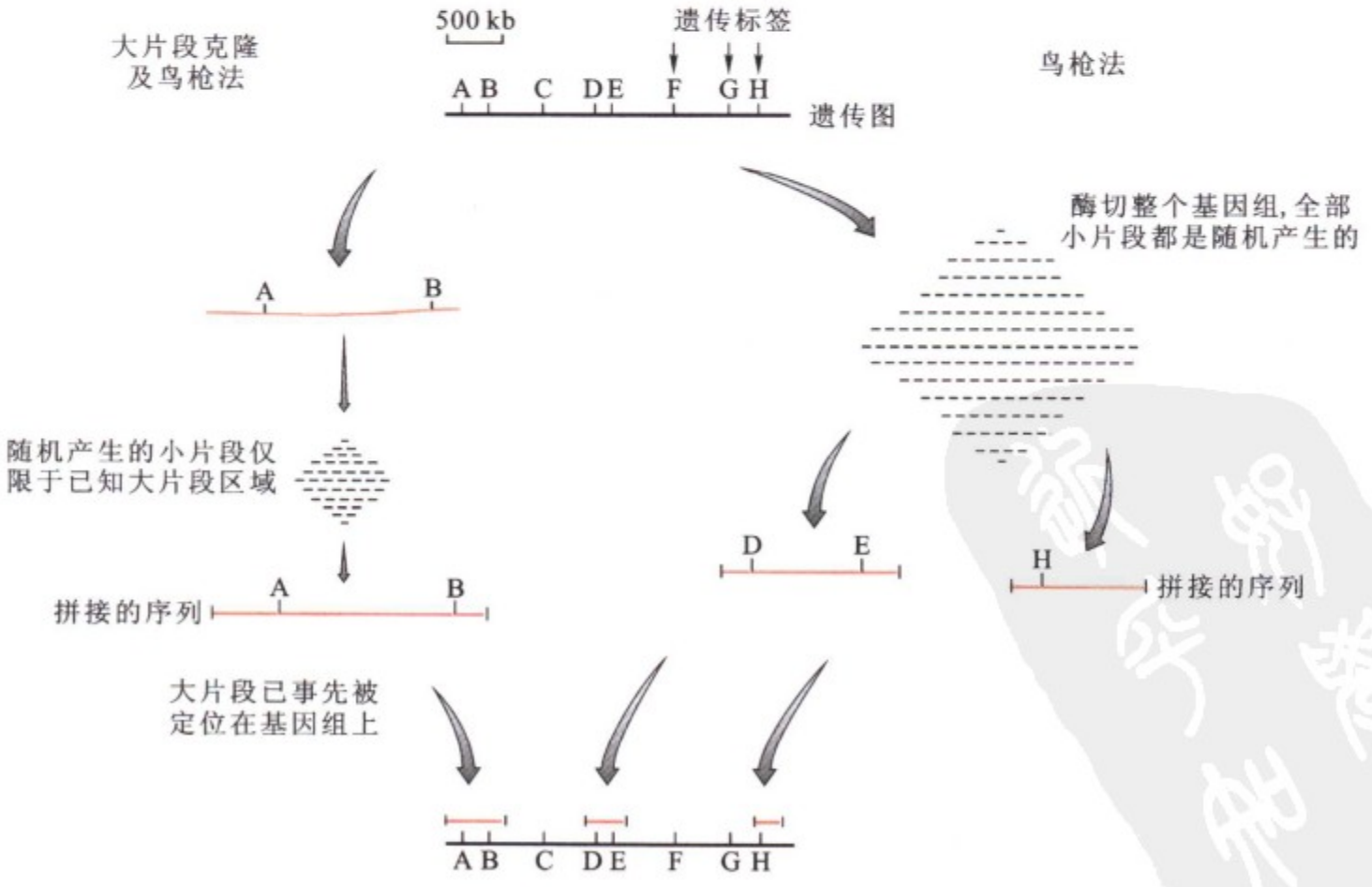



图 11-6 改进后的鸟枪法测序原理图

### 11.1.4 大规模 DNA 序列拼接

目前,对于 DNA 测序的技术只能一次测序大约 1 kb 左右的 DNA 序列,可是人类的 DNA 全长大约为 30 亿个, 大约是 300 万个 1 kb 的集合。如果 DNA 测序能够沿模板次序进行,一次 1 kb 可能也不是问题,只不过重复的次数多了些。事实上,每次测序是在染色体 DNA 中随机取一小段,因此,把这些随机的小片段按原样拼接产生完整的 DNA 序列就成为基因组序列分析中亟需解决的最关键的问题。

DNA 序列拼接问题与组合数学中的最短超串问题相似。最短超串问题

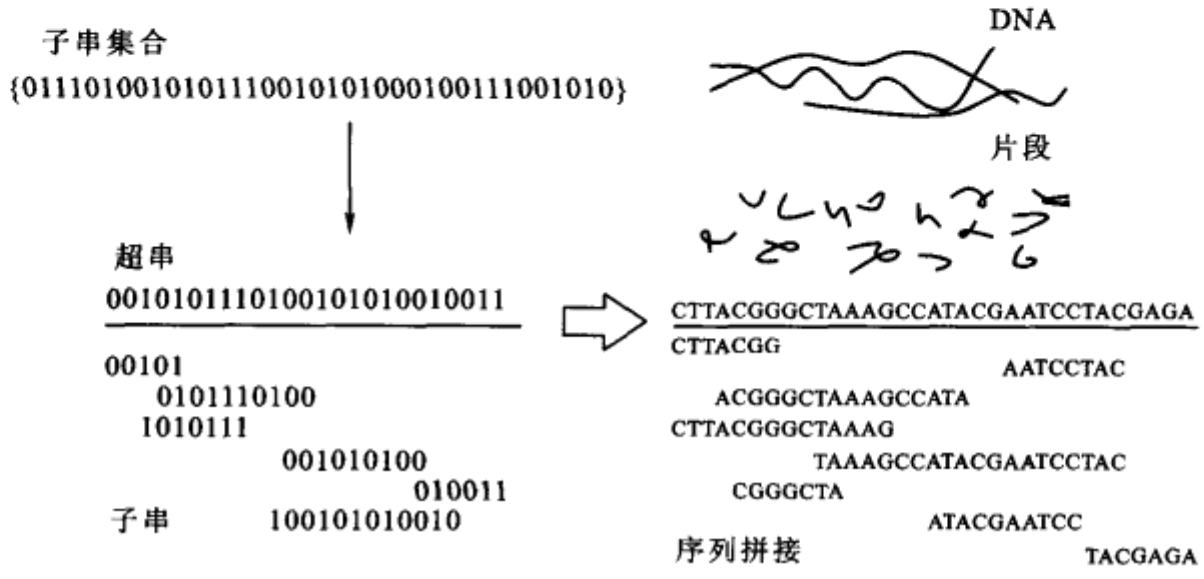



**图 11-7 数学中的最短超串问题与 DNA 的序列拼接**

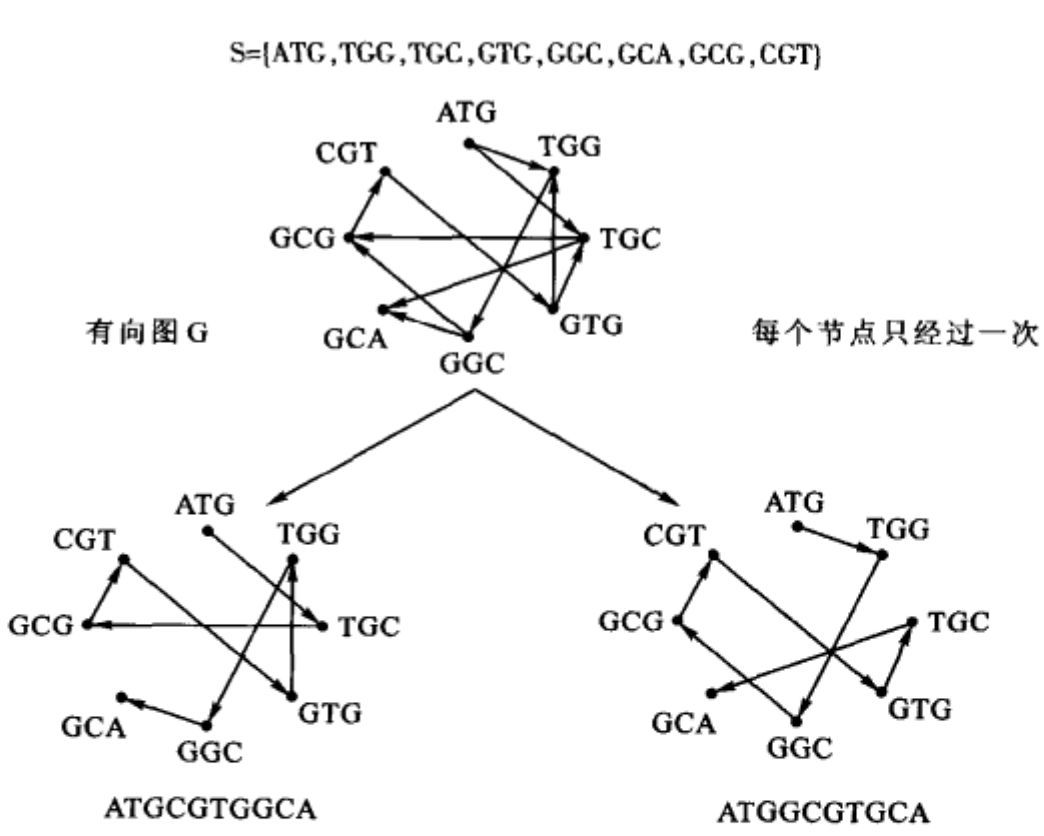



**图 11-8 汉米尔顿通路法** 首先将全部数据构成一个有向图 G,将每个数据看成一个节点,如果两个数据之间有重叠(overlap),那么,通过寻找经过每个数据一次且仅一次的一条路径就可以将序列拼接问题转化成汉米尔顿通路问题。

即给定一个字符串的集合，找出一最短的字符串称为超串，并将集合中的任何一元素作为其子串(图 11-7)。在差值=0 的情况下，DNA 序列拼接问题实质上就是最短超串问题。由于序列拼接问题的复杂性以及在生物信息学中的特殊重要性，人们对这一问题给予了极大的重视，并提出了不少算法。目前较为通行的主要有两类，即汉米尔顿通路法(Hamiltonian path)和尤路因通路法(Eulerian path)。两种方法都是基于图论的近似算法，首先根据片段间的重叠信息构造某种类型的图，然后利用图论的计算方法获得实验结果(图 11-8，11-9)。

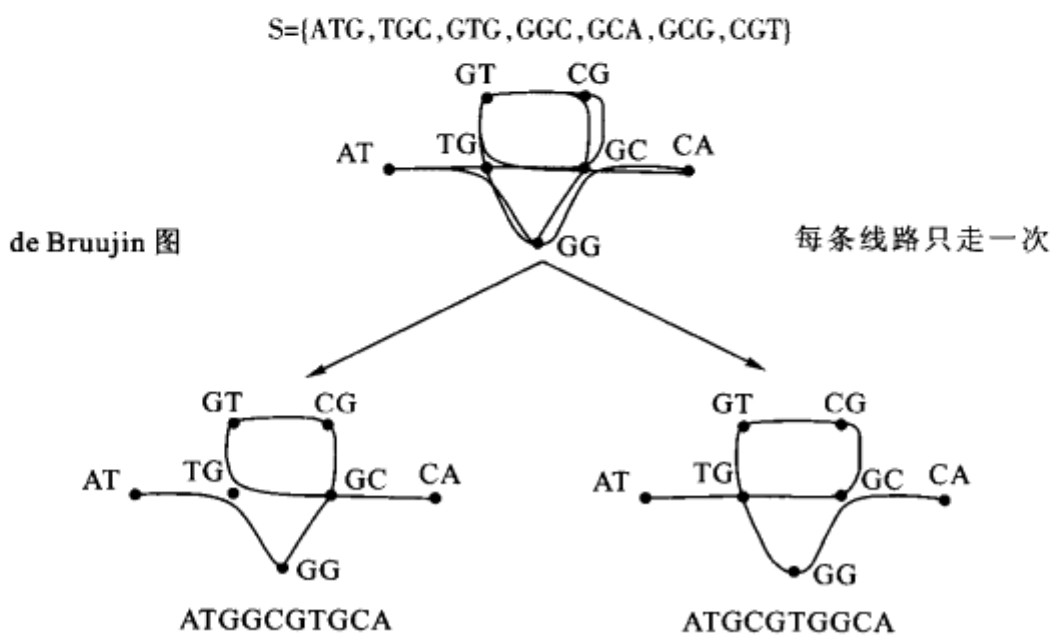



**图 11-9 尤路因通路法** EULER 将序列拼接归纳为寻找超级通路的问题，它不研究各个数据之间的重叠而是将各个数据切割成规整的小片段，并按照以下三步进行运算：①将所有数据切割成小片段(即 k-子串)，以排除数据中的错误，获得准确度较高的次级数据；②构建 de Bruujin 图，每个次级数据表示成一条通路，每个重复表示一个多入口、多出口的组合，但不知道出入口之间的对应关系；③通过等价变换，在 de Bruujin 图中寻找一条尤路因通路，使之能覆盖所有的通路。

## 11.2 人类基因组计划

基因组学这一名词是美国人 T.H.Roderick 在 1986 年 7 月造出来的，与一个新的杂志——Genomics 一道问世。基因组学完全改变了经典遗传学“零敲碎打”的方法，因为它着眼于研究并解析生物体整个基因组的所有遗传信息。基因组是生物体内遗传信息的集合，是某个特定物种细胞内全部 DNA 分子的总和。人类基因组包括 23 对染色体，单倍体细胞中约有 30 亿对核苷酸，编码了 2 万~3 万个基因，人类基因组中携带了有关人类个体生长发育、生老病死的全部遗传信息。从整体上看，不同人类个体的基因是相同的，因此，我们说“人类只有一个基因组”，人生来是平等的。当然，不同的人可能拥有不同的等位基因，这一点决定了人与人之间个体上的差异。

### 11.2.1 人类基因组计划的科学意义

确定人类基因组所携带的全部遗传信息,认识自我,揭开人类生长发育的奥秘,追求健康,战胜疾病,是人类基因组计划的最终目标,也是人类自然科学史上的新篇章。图 11-10 总结了包括人类基因组在内的主要基因组计划的基本情况。到目前为止,已经完成了酵母、线虫、果蝇、拟南芥和人类 5 个真核生物基因组以及大肠杆菌等数十个原核生物基因组。

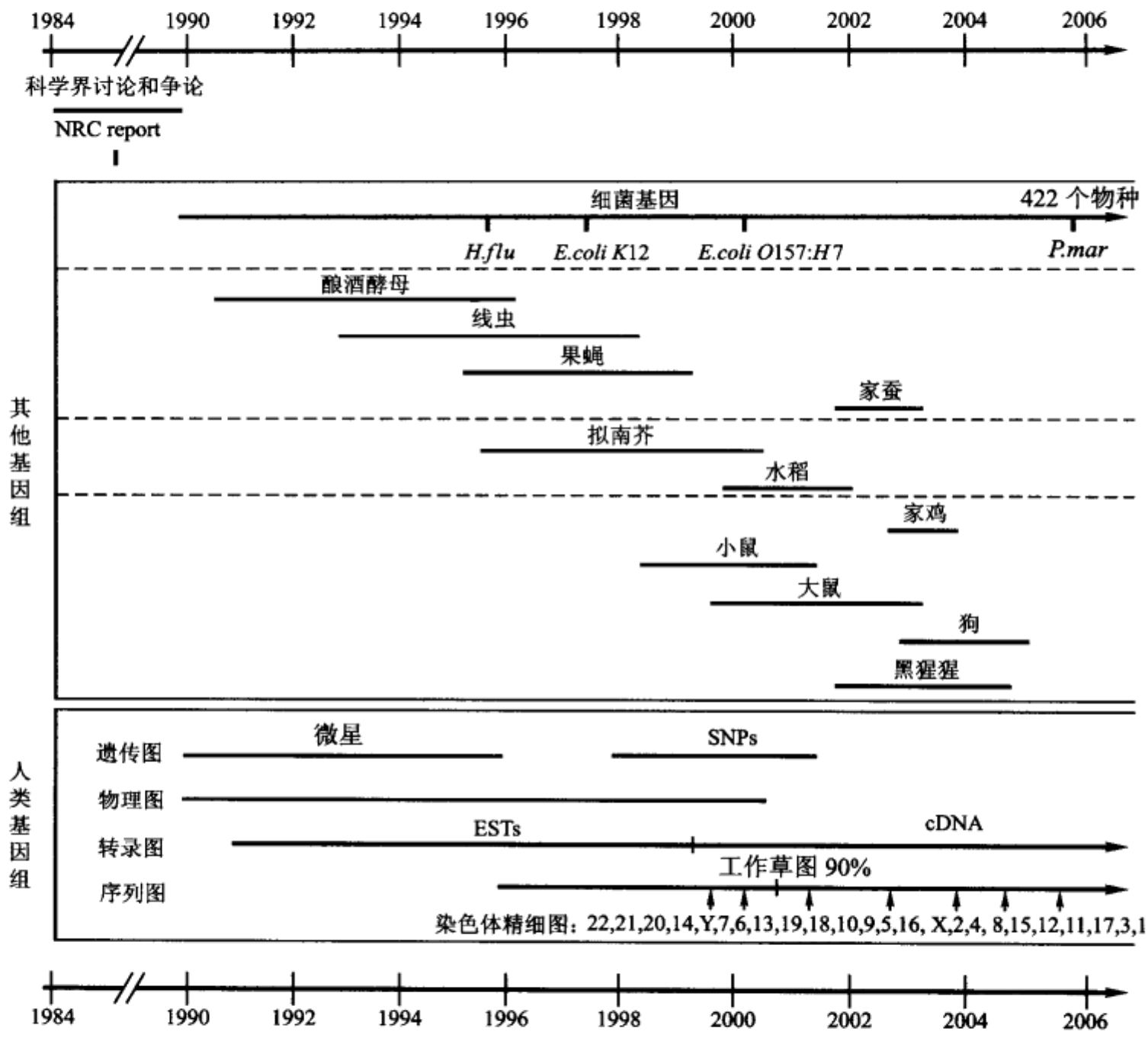



图 11-10 到 2006 年 12 月全世界主要基因组计划的进展情况

人类基因组计划的科学意义在于:

(1) 确定人类基因组中 3 万个左右编码基因的序列及其在基因组中的物理位置,研究基因的产物及其功能。

(2) 了解转录和剪接调控元件的结构与位置,从整个基因组结构的宏观水平上理解基因转录与转录后调节。

(3) 从整体上了解染色体结构,包括各种重复序列以及非转录“框架序列”的大小和组织,了解各种不同序列在形成染色体结构、DNA 复制、基因转录及表达调控中的影响与作用。

(4) 研究空间结构对基因调节的作用。有些基因的表达调控序列与被调节基因从直线距离上看,似乎相距甚远,但若从整个染色体的空间结构上看则恰恰处于最佳的调节位置,因此,有必要从三维空间的角度来研究真核基因的表达调控规律。

(5) 发现与 DNA 复制、重组等有关的序列。DNA 的忠实复制保障了遗传的稳定性,正常的重组提供了变异与进化的分子基础。局部 DNA 的推迟复制、异常重组等现象则导致疾病或者胚胎不能正常发育,因此,了解与人类 DNA 正常复制和重组有关的序列及其变化,将对研究人类基因组的遗传与进化提供重要的结构上的依据。

(6) 研究 DNA 突变、重排和染色体断裂等,了解疾病的分子机制,包括遗传性疾病、易感性疾病、放射性疾病甚至感染性疾病引发的分子病理学改变及其进程,为这些疾病的诊断、预防和治疗提供理论依据。

(7) 确定人类基因组中转座子、反转座子和病毒残余序列,研究其周围序列的性质。了解有关病毒基因组侵染人类基因组后的影响,可能指导人类有效地利用病毒载体进行基因治疗。

(8) 研究染色体和个体之间的多态性。这些知识可被广泛用于基因诊断、个体识别、亲子鉴定、组织配型、发育进化等许多医疗、司法和人类学的研究。此外,这些遗传信息还有助于研究人类历史进程、人类在地球上的分布与迁移以及人类与其他物种之间的比较。

### 11.2.2 遗传图

遗传图(genetic map)又称为连锁图(linkage map),是指基因或 DNA 标志在染色体上的相对位置与遗传距离,后者通常以基因或 DNA 片段在染色体交换过程中的分离频率厘摩(cM)来表示。cM 值越大,两者之间距离越远。通过遗传图分析,我们可以大致了解各个基因或 DNA 片段之间的相对距离与方向,了解哪个基因更靠近着丝粒,哪个更靠近端粒等。遗传距离是通过遗传连锁分析获得的,研究中所使用的 DNA 标志越多,越密集,所得到的遗传连锁图的分辨率就越高。

连锁分析是经典遗传学的重要内容,其实质是通过分析同一遗传位点在不同个体中等位基因的不同(多态性)来研究同一染色体上两个位点之间的相互关系。因为在产生配子的减数分裂过程中,亲代同“号”的父源或母源染色体既能相互配对也可能发生片段互换,而父母源染色体等位基因互换导致子代出现 DNA“重组”的频率与这两个位点之间的距离呈正相关。所以,科学上用两个位点之间的交换或重组频率来表示其“遗传学距离”(图 11-11)。一般说来,这一数值不会大于 50%或 50 cM,因为当重组率等于 50%(即遗传学距离等于 50 cM)时,两个位点之间完全不连锁,发生随机交换。

**图 11-11 遗传距离图的基本数据来自基因的重组** 注：上述 4 个基因都位于果蝇的 X 染色体上。

经典的遗传标记是可被电泳或免疫技术检出的蛋白质标记，如红细胞 ABO 血型位点标记，白细胞 HLA 位点标记等。表 11-1 给出了酵母遗传分析中最常见的部分生物化学标签。由于人类本身不能像其他“非人类”生物那样进行“选择性”婚配，子代个体数量较为有限、世代寿命较长等客观原因，已知呈共显多态性的蛋白质数量不多，等位基因的数量（多态性）也不多，使人类遗传性状研究受到很大的限制。DNA 技术的建立为人类提供了大量新的遗传标记。

**表 11-1 酵母遗传分析中最常用的生物化学标签**

| 标签 | 表 现 型 | 筛 选 方 法 |
|---|---|---|
| ADE2 | 培养基中需加入腺苷酸 | 只能在加入腺苷酸的培养基上生长 |
| CAN1 | 对刀豆氨酸有抗性 | 能在含有刀豆氨酸的培养基上生长 |
| CUP1 | 对铜离子有抗性 | 能在含有铜离子的培养基上生长 |
| CYH1 | 对环己酰亚胺有抗性 | 能在含有环己酰亚胺的培养基上生长 |
| LEU2 | 培养基中需加入亮氨酸 | 只能在加入亮氨酸的培养基上生长 |
| SUC2 | 能进行蔗糖发酵 | 能在以蔗糖作为唯一碳源的培养基上生长 |
| URA3 | 培养基中需加入尿嘧啶 | 只能在加入尿嘧啶的培养基上生长 |

第一代 DNA 遗传标记是 RFLP（restriction fragment length polymorphism，限制性片段长度多态性）。DNA 序列上的微小变化，甚至 1 个核苷酸的变化，也能引起限制性内切酶切点的丢失或产生，导致酶切片段长度的变化（图 11-12）。由于核苷酸序列的改变遍及整个基因组，特别是进化中选择压力不是很大的非编码序列之中，RFLP 的出现频率远远超过了经典的蛋白质多态性。而且，只要

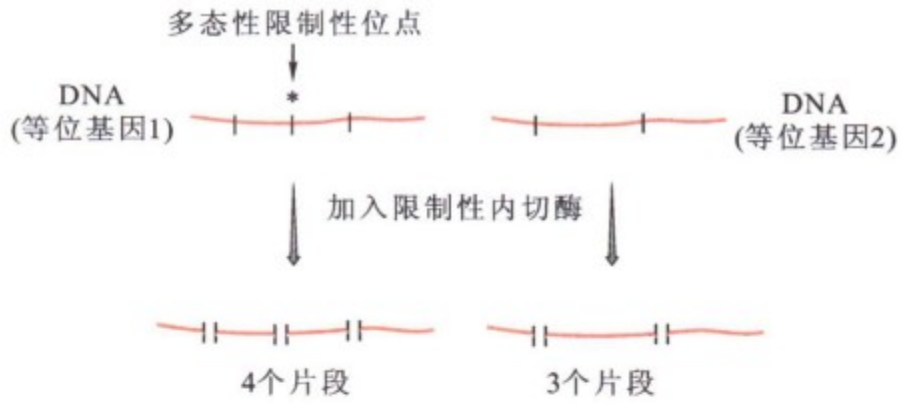



图 11-12 限制性片段长度多态性(RFLP)原理示意图

选择得当，生物体内出现共显性 RFLP 及 RAPD 分子标记的频率较高（图 11-13）。但是，RFLP 仅局限于 1 个或少数几个核苷酸的突变，一般只能产生限制性酶切位点的“切开”与“不切开”两种情况，所提供的“多态性”信息量较小。进行

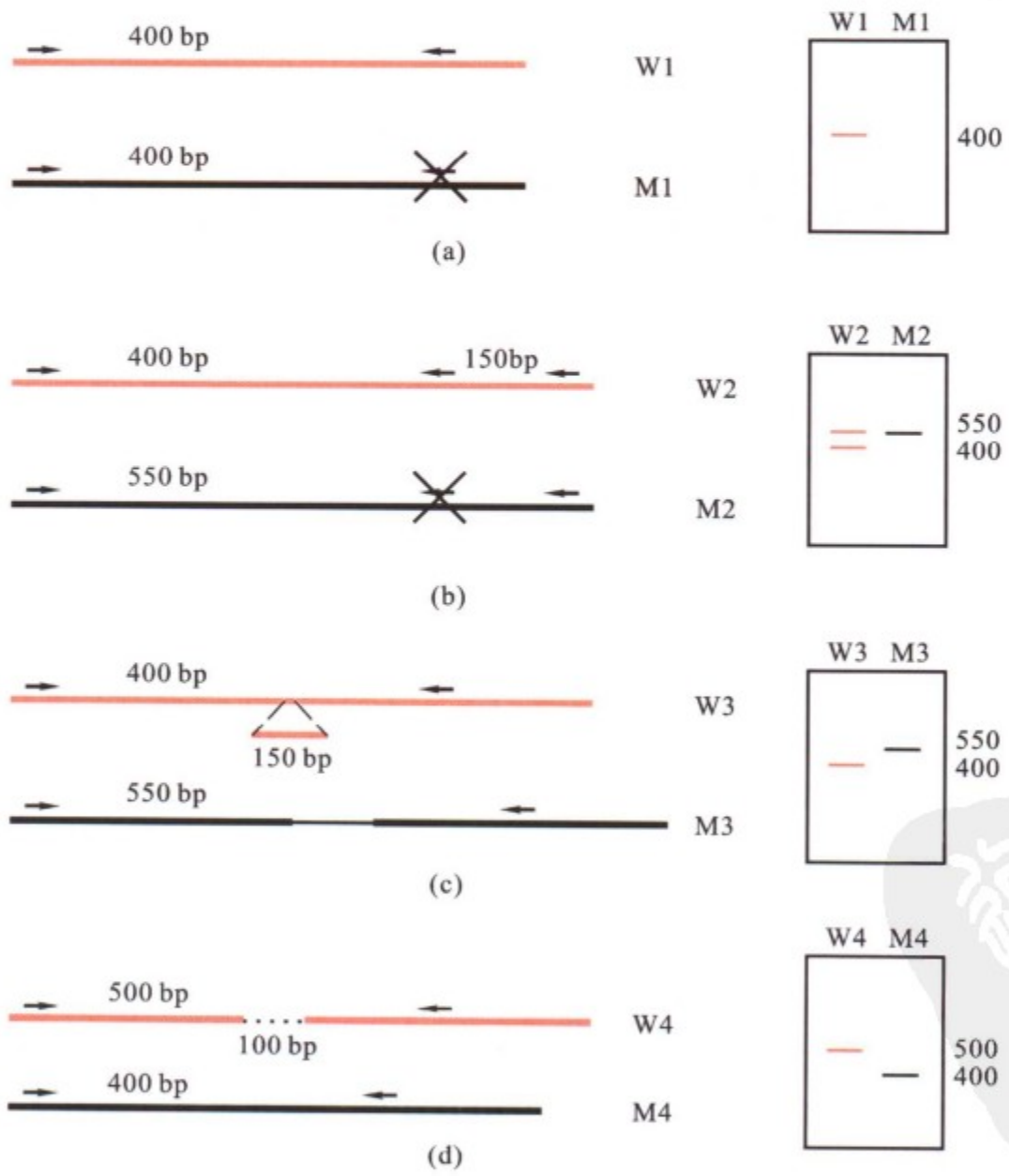



图 11-13 RFLP 分子标记中的显性与共显多态性分子机制 (a)引物结合位点突变-1；(b)引物结合位点突变-2；(c)插入突变；(d)缺失突变 (a)是最常见的显性多态性标记，(b)，(c)，(d)都是共显性标记。

RFLP 分析时,需要该位点的 DNA 片段做探针,用放射性同位素及核酸杂交技术,既不安全又不易自动化。

第二代 DNA 遗传标记利用了存在于人类基因组中的大量重复序列,包括重复单位长度在 15~65 个核苷酸左右的小卫星 DNA(minisatellite DNA),重复单位长度在 2~6 个核苷酸之间的微卫星 DNA(microsatellite DNA),后者又称为简短串联重复 (STR、STRP 或 SSLP,short tandem repeat polymorphism 或者 simple sequence length polymorphism)。STRP 有两个最突出的优点,即作为遗传标记的"多态性"与"高频率"。由于 $A_n$,$(CA)_n$,$(CGG)_n$ 等短重复序列在进化上不受选择,在同一位点上可重复单位数量变化很大,同源染色体配对时又容易产生"错配",进一步扩大了群体中所拥有的"等位基因"数,并使这样的位点遍布于整个基因组(表 11-2)。已有 5 264 个 STRP 为主体的遗传标记"连锁图",平均分辨率已达 600 kb,其中第 17 号染色体上平均每 495 kb 有一个标记,第 9 号染色体上平均每 767 kb 有一个标记,整个基因组中只有三处标记间距大于 4 Mb(图 11-14)。

表 11-2 人类基因组中的各种主要卫星 DNA 比较

| 卫星 DNA 分类 | 特 征 |
|---|---|
| α 卫星 DNA | 中等重复,基本单位长 171 bp |
| 小卫星 DNA | 中等重复,基本单位长 15 ~ 65 bp |
| 微卫星 DNA | 中等重复,基本单位长 2 ~ 8 bp |

卫星 DNA

串联重复的基本单位首尾相接,在基因组中呈不均匀分布,但主要集中于着丝粒、端粒等特定部位,高度或中等重复,分属三个大家族。

人类基因组中的遗传多态性大多表现在重复序列上,特别是短串联重复序列,如小卫星 DNA 和微卫星 DNA 上,其多态性主要来自重复序列拷贝数的变化。小卫星 DNA 的长度一般不超过 20 kb,由 15~65 bp 的基本单位串联重复而成。小卫星 DNA 散布于人类各染色体的不同部位,有的位于基因内部,也有的位于基因之外,其重复次数(即小卫星 DNA 区的长度)在人群中是高度变异的。小卫星 DNA 有高度变异性但却按照孟德尔的规律遗传,所以,它们是很好的 DNA 多态性标记,被广泛应用于基因定位、DNA 指纹分析和遗传病的连锁诊断。

与小卫星 DNA 相类似,微卫星 DNA 也是由重复序列串联而成的,它们的基本序列只有 2~8 bp,如 $(TA)_n$ 及 $(CGG)_n$ 等,通常只重复 10~60 次。人类基因组中已定位的微卫星位点超过 10 000 处,因其重复次数在不同个体间呈高度变异性而被广泛应用于遗传分析和疾病诊断中。可用酶切后电泳并进行杂交的方法检测,也可用 PCR 扩增后直接观察电泳条带迁移率变化来判断。

占人类基因组约 45%的重复序列来源于转座子复制机制(重复序列共占人类基因组序列总量的 50%以上)。序列分析表明,四类转座子产生了这些重

图 11-14 人类基因组微卫星遗传标记图

复序列(图 11-15),其中前三类转座子以 RNA 为中间产物进行转座,最后一类则直接以 DNA 的形式转座。LINEs(long interspersed elements)可能是人类基因组中发现的最古老的重复序列,这些转座单位一般长 5~6 kb,含有 RNA 聚合酶Ⅱ启动子序列和两个可译框架,一旦被翻译,一个 RNA 与自身编码的

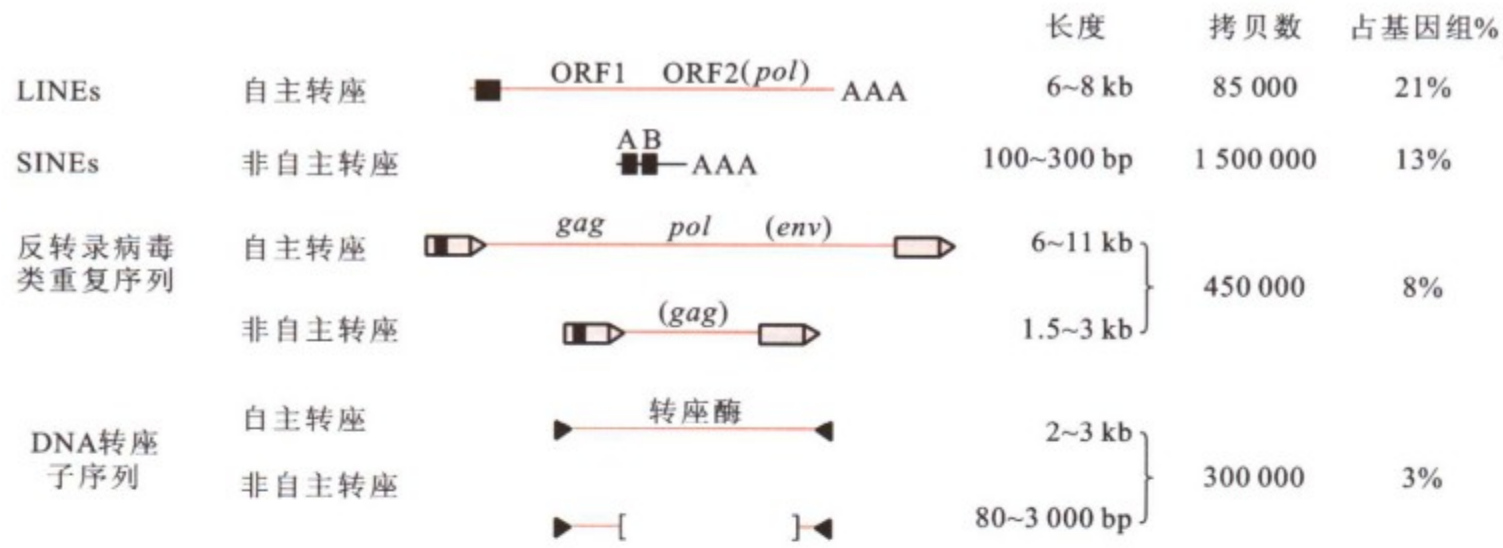


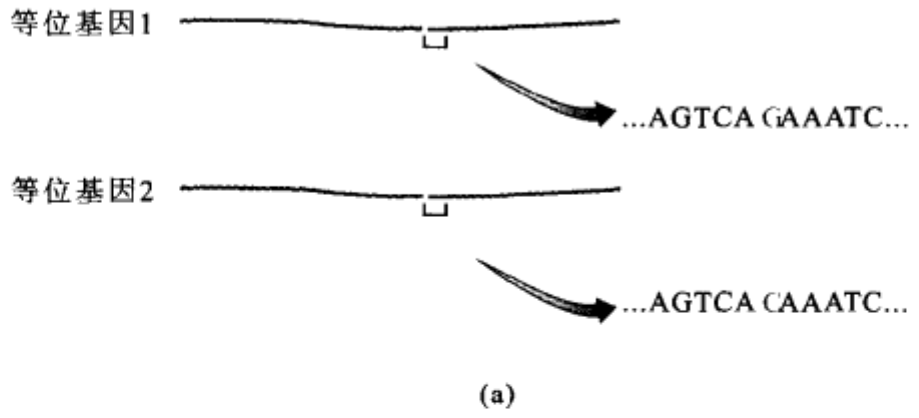

图 11-15 存在于人类基因组重复序列中的四类转座子序列分析

蛋白质结合后被运转到核内,并从 3′端开始反转录,产生许多没有完整 5′端的 DNA 片段。因此,由 LINE 转座子产生的重复单元大多是不完整的,现在仍然有转座活性的LINE1 产生的重复单元平均长度为 1 070 bp,新生 DNA 片段两端有 7~20 bp 的重复序列。SINEs(short interspersed elements)是非自主转座子,长度在 100~400 bp,其 3′端与 LINEs 有同源性,因此能依靠 LINEs 进行转座。

第三代 DNA 遗传标记,可能也是最好的遗传标记,是分散于基因组中的单个碱基的差异。这种差异包括单个碱基的缺失和插入,但更常见的是单个核苷酸的替换,即单核苷酸的多态性(SNP,single nucleotide polymorphism)。作为一种碱基的替换,SNP 中大多数为转换,即由一种嘧啶碱基替换另一种嘧啶碱基,或由一种嘌呤碱基替换另一种嘌呤碱基。相应地,由嘧啶碱基替换原有的嘌呤碱基或由嘌呤碱基替换原有的嘧啶碱基就称为颠换。颠换与转换之比为 1∶2。SNP 在 CG 序列上出现最为频繁,而且大多是 C→T,因为 CG 中的 C 即胞嘧啶常被甲基化,自发脱氨后变成胸腺嘧啶(T)。

由于该标记中的所有“遗传多态性”都来自单个核苷酸的差异,SNP 有可能在密度上达到人类基因组“多态”位点数目的极限。到目前为止已经在人类基因组发现了超过 1 000 万个 SNP 位点,平均每 300 bp 中就有一个 SNP! 由于遗传中的选择压力,也由于基因组中蛋白质编码的序列仅占 10%以下,绝大多数 SNP 位于非编码区。SNP 与 RFLP 和 STRP 标记的主要不同之处在于,它不再以 DNA 片段的长度变化作为检测手段,而直接以序列变异作为标记(图 11-16)。SNP 遗传标记分析完全摒弃了经典的凝胶电泳,代之以最新的 DNA 芯片技术,可能是人类基因组“遗传图”发展方向。

“遗传图”的建立为人类疾病相关基因的分离克隆奠定了基础。拥有 5 000 多个遗传学位点,相当于把整个人类基因组划分为 5 000 多个小区,并分别设置了“标牌”。这些标牌将在搜索功能基因的过程中发挥独特的作用。把多态性的疾病基因位点(该位点至少包括“正常”及“致病”两个等位基因)与上述遗传标记进行分析比较时,如果在家系中证实该基因与某个标记不连锁(重组率为 50%),表明该基因不在这一标记附近;如果发现该基因与某个标记有一定程度




▌图 11-16 人类基因组中的 SNP 作为遗传标记的分子机制 (a)单核苷酸多态性的产生;(b)SNP 的产生影响了 DNA 序列间杂交的强度。

的“连锁”(重组率小于 50%但大于 0),表明它可能位于这个标记附近;如果该基因与某标记间不发生重组(重组率等于 0),我们就推测该标记与所研究的疾病基因可能非常接近。

### 11.2.3 物理图

人类基因组的物理图 (physical map) 是指以已知核苷酸序列的 DNA 片段(序列标签位点,sequence-tagged site,STS)为“路标”,以碱基对(bp,kb,Mb)作为基本测量单位(图距)的基因组图。任何 DNA 序列,只要知道它在基因组中的位置,都能被用作 STS 标签。物理图的主要内容是建立相互重叠连接的“相连DNA片段群”(contigs),并用 PCR 方法予以证实。图 11-17是酵母第三号染色体的遗传图与物理图的比较。序列分析表明,遗传图上关于 *glk1* 和 *cha1* 两个基因位点的定位是错误的。由于实验方法不同,不少遗传标记之间的距离并不等于它们在物理图上距离。

用 STS 技术绘制基因组物理图是到目前为止最为有效的方法。STS 其实只是基因组中任何单拷贝的短 DNA 序列,长度在 100~500 bp。建立 STS 物理图之

果常常不能预测；另外还有一个问题，即很多人们感兴趣的蛋白，特别是膜蛋白疏水性很高，很难溶于用于双向电泳的缓冲液中，电泳出来根本不可见；此外，很多蛋白质的含量太低，双向电泳检测不到。目前上述问题还难以解决，但研究人员对不同细胞成分进行分离再进行研究，在一定程度上解决了这些问题。例如，可以只研究核蛋白，甚至只研究核仁或是核孔复合物的蛋白。需要分离的蛋白质更少，溶解就不会成为一个严重问题。不过，这些受限制的研究离真正的蛋白质组学研究还很遥远。

**2. 蛋白质分析**

一旦蛋白质可以被分离和定量，如何进行分析？首先，它们要得到鉴定，目前最好的方法是：将单个蛋白质斑点切出，用蛋白酶将其裂解成肽段，这些肽段再用质谱仪进行确定。图 19-11 显示普遍使用的基质辅助激光解析电离 - 飞行质谱（Matrix-Assisted Laser Desorption Ionization time-of-flight，MALDI - TOF）技术：肽段首先置于基质上，形成晶体，利用激光束将晶体离子化，然后加大电压使离子化的晶体飞向检测器。一般晶体离子化只是带上一个电荷，因此飞行时间取决于晶体的质量，质量越大，飞行的时间越长。在 MALDI - TOF 中，离子束还可以通过静电反射器发生偏转，这样通过第二个检测仪就可以更精确确定晶体的质量，而这些质量数据可以确切地反映肽段的化学组成。如果整个基因组的序列已知，清楚蛋白质的编码情况，那么就可以运用计算机技术分析质谱结果，可以将双向电泳的每个斑点与基因组中的一个基因联系起来，进而预测蛋白质的序列。不过，蛋白质序列的测定并不代表对蛋白质活性的认识，很多蛋白质的性质特点还需要进一步研究确定。

与 DNA 芯片一样，可以发展蛋白质芯片，用于快速鉴定几千个蛋白质，这样就可以省去了蛋白质分离步骤。蛋白质芯片的一个策略是产生能特异和定量识别蛋白质的抗体并将它们结合在芯片上。蛋白质芯片技术的成熟还有很多障碍要跨越，首先是抗体的制备，这比寡核苷酸的制备要费时和昂贵得多，事实上，就目前技术而言，要获得每个蛋白质的特异性抗体还不可能；并且，对于那些用双向电泳都检测不出的低丰度蛋白，要用更小型化的芯片进行检测，难度更大。人类基因组序列的测定在相当长时间内进展不大，但是 20 世纪 80 年代，人类基因组计划的提出极大地促进了该计划的完成，也许有关蛋白质组的研究也需要一个全面的蛋白质组计划来促进。

一个生物体产生的所有 RNA 称为转录组，而对这些转录产物进行的研究称为转录组学。类似地，一个生物体产生的所有蛋白质称为蛋白质组，相应的研究称为蛋白质组学。蛋白质组学首先要求在广泛的范围内将蛋白质分离，目前最好的方法是蛋白质双向电泳，其次是蛋白质的鉴定。目前最好的方法是将蛋白质用蛋白质酶酶解后，用质谱仪进行分析。在未来，抗体芯片可能用于蛋白质分析，蛋白质分离这一步骤可以省去。

了基因组的主要成分（可转录部分）。人类的基因转录图（expression profiling）（cDNA 图），或者基因的 cDNA 片段图，即表达序列标签图（expressed sequence tag，EST）是人类基因组图的雏形。整个人类基因组中，只有 1%~5%的序列编码了蛋白质，最多可能有 5~7 万个蛋白质编码基因。表 11-3 是几种已经完成 DNA 序列分析生物细胞内蛋白质种类的比较，可以看出原核生物、低等真核生物和高等真核生物之间存在很大的差异。在成年个体的每一特定组织中，一般只有 10%~20%的结构基因（约 1~2 万个不同类型的 mRNA）表达。从研究的角度看，得到了一段 cDNA 或一个 EST，就能被用于筛选全长的转录本，并将该基因准确地定位于基因组上，所以，转录图是基因组计划的重要组成部分。

表 11-3 不同生物基因组的蛋白质编码能力比较

| 物 种 | 基因组总长度 | 蛋白质种数/个 |
|---|---|---|
| 尿殖道支原体 *Mycoplasma genitalium* | 580 073 bp | 467 |
| 肺炎支原体 *Mycoplasma pneumoniae* | 816 394 bp | 677 |
| 流感嗜血杆菌 *Haemophilus influenzae* | 1 830 138 bp | 1 709 |
| 枯草芽胞杆菌 *Bacillus subtilis* | 4 214 814 bp | 4 100 |
| 大肠杆菌 *Escherichia coli* | 4 639 221 bp | 4 288 |
| 酿酒酵母 *Saccharomyces cerevisiae* | 13 116 818 bp | 6 275 |
| 线虫 *Caenorhabditis elegans* | 约 97 Mb | 18 891 |
| 拟南芥 *Arabidopsis thaliana* | 115 Mb | 25 498 |
| 果蝇 *Drosophila melanogaster* | 116 Mb | 14 113 |
| 人类 *Homo sapiens* | $3.2 \times 10^9$ bp | 约 2.5 万左右 |

大规模生产 EST 的主要程序如下：分离特定组织在某一发展阶段或某种生理条件下的总 mRNA，合成 cDNA 并进行序列分析。任何一段长度达到几百个核苷酸的 cDNA 序列都具有转录本的特异性，代表了不同基因的信息。一般说来，常将基因表达谱中的已知基因按照其功能和亚细胞定位分成若干类，主要包括蛋白质合成相关蛋白、细胞骨架蛋白、细胞质蛋白、核蛋白、膜蛋白、分泌蛋白、未定位或功能未知蛋白。研究中，往往再根据不同的细胞表型把各类蛋白分成若干亚类，如胞质蛋白就被分为与能量代谢相关蛋白、溶酶体蛋白、信号转导相关蛋白等。

收集各种细胞或组织的基因表达谱进行两两或多重比较，即所谓的“电子消减杂交”或“数据库消减杂交”，就能较全面地了解哪些基因是特异性表达的、哪些是上调的、哪些是下调的，哪些是持家基因。在某一细胞或组织中特异性表达的基因可能与该组织或细胞类型的生理功能有关，这是表达谱比较研究的理论基础。收集各类组织或细胞的基因表达谱并把每个基因都标记到所表达的组织中，就可以绘出由 200 余种人体基本组织或不同细

胞组成的人体基因图(body-map)。转录图或基因表达谱研究所提供的信息,使人们有可能系统、全面地从 mRNA 水平了解特定细胞、组织或器官的基因表达模式并解释其生理属性,深入认识细胞生长、发育、分化、衰老和疾病发生的机制。

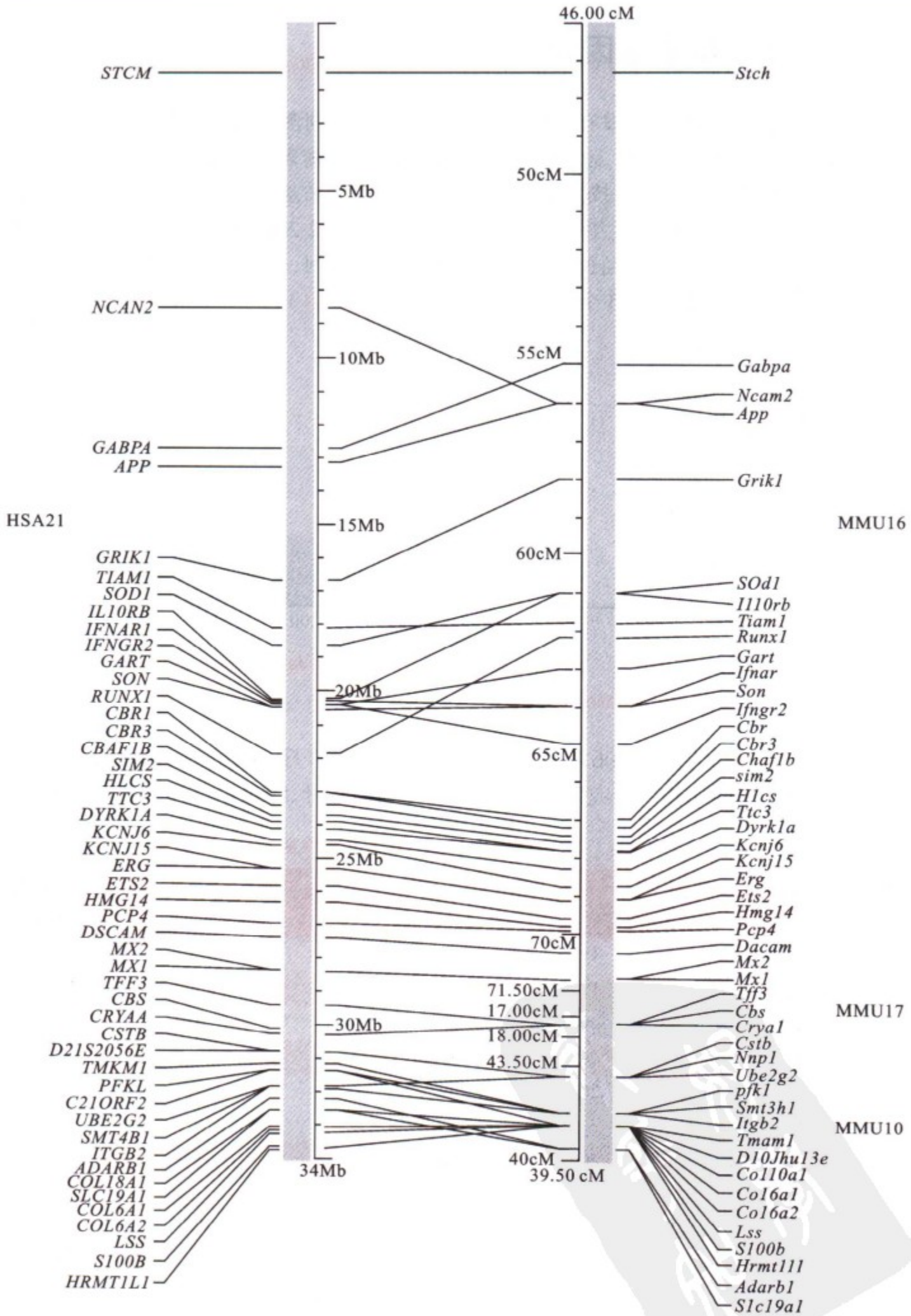
### 11.2.5 全序列图

人类基因组的核苷酸序列图(human genome sequence)其实是分子水平上最高层次的、最详尽的物理图。测定总长约 1 m、由 30 亿个核苷酸组成的全序列当然是人类基因组计划中最明确、最艰巨的任务(图 11-19)。因为人类所拥有的基因位点都是相同的,不同种族、不同个体的基因差异(人类基因组的多样性)以及"正常"与"疾病"基因的差异,只是同一位点上的等位基因的差异,所以,现在的人类基因组全序列来自一个"代表性人类个体"(其所有权在法律上不属于任何供体)。该序列在理论上代表了全人类的基因组信息,事实上也可被用于任何族群、任何个体的基因分析和诊断。也正因为如此,一个人类基因的专利,无论其实验材料来自何地何人,都意味着全人类的这一基因被专利,不存在白人或黑人、外国人或中国人对该基因的争议,尽管白人与黑人、外国人与中国人在这个基因位点上必然存在某种变异。研究还发现,人类基因组与其他动物基因组在染色体水平上有"共线"(即同源)现象。图11-20是人类第 21 号染色体 HSA21 位点与小鼠第 16 号染色体 MMU16,MMU17 和 MMU10 连锁图的比较,我们可以很清楚地看出,两者之间存在着广泛的同源性。

人类基因组计划所提供的这四张图,特别是人类核酸序列图,蕴藏了决定我们生、老、病、死的所有遗传信息,必将成为人类认识自我、改造自我的取之不尽的知识源泉,为 21 世纪现代生物学和医学的迅速发展奠定基础。

cctattagaa cgattaagga gaaatacaat tcgaatgaga aggatgtgcc gtttgttata ataaacagcc acacgacgta aacgtaaaat gaccacatga tgggccaata gacatggacc gactactaat aatagtaagt tacattttag gatggaataa atatcatacc gacatcaggtt tgaaagaaaa gggaaaaaaa gaaaaaataa ataaaagata tactaccgac atgagttcca aaaagcaaaa aaaaagatca agccgacaca gacacgcgta gagagcaaaa tgactttgac gtcacaccac gaaaacagac gcttcatacg tgtcccttta tctctctcag tctctc tataaa cttagtga gaccctcctc tgtttt actcAca aatatgc aaactagaaa acaatcatca ggaataaagg gtttgattac ttctattgga aagaaaaaaa tctttggaaa ATGGATCAAACAGAGGAACCACCACTCAACACACACCAGCAGCACCCAGGTAGATT CTAATTTCAGAAACTTATATTTTTTTTAAGTGACAATCCTCTGAATTTACTTAAACTTATTGT GATTTATGGATACAGAAGAAGTTGAACATCATGAGAATGGTGCGACTAAGATGTTTAGG AAAGTTAAAGGCTAGAGCTAAGAAGTTCAAGAACAGTCTCACTAAACATGGACAAAGC

**图 11-19 真核基因代表性序列结构示意图** 拟南芥(*Arabidopsis thaliana*)基因 *RD29A*(Locus ID:At5g52310)序列:方框中的黑色大写字母表示编码区(外显子,exon),方框之间的灰色大写字母表示非编码区(内含子,intron),方框中第一个 ATG 表示翻译起始位点。红色字母分别表示 CAAT 区,TATA 区及转录起始位点序列 actcAca,其中大写的 A 是转录起点,一般记作+1 位。向上游记作负值(-1,-2,-3…),向下游记作正值(+1,+2,+3…)。




▌图 11–20 人第 21 号染色体 HSA21 位点与小鼠 16 号染色体 MMU16,MMU17 和 MMU10 位点有“共线”性

## 11.3 其他基因组

若干年前,几乎没有人相信有一天基因组序列测定的增长速度会超过著名的摩尔定律关于 CPU 运算能力的定义。事实上,随着焦磷酸测序法等新技术的逐步成熟和现有测序技术成本的持续降低,近年来基因组序列数据每年呈指数形式增长。根据 2007 年 1 月的数据,全球已启动 2 296 个基因组项目,其中 607 个项目已经完成,已经公开发表 481 个基因组序列,包括 403 个细菌基因组,33 个古细菌基因组和 45 个真核生物基因组(图 11-21)。

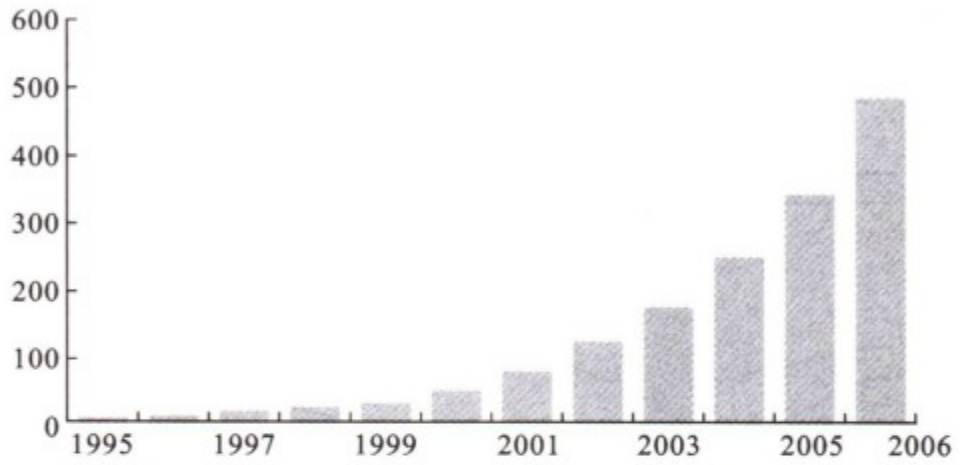



图 11-21 到 2006 年年底已完成的基因组项目(http://www.genomesonline.org/)

表 11-4 是到 2006 年为止已经基本完成 DNA 序列分析的各种真核生物基因组数据的比较。研究发现,低等真核生物如酵母、线虫以及高等植物拟南芥,

表 11-4 基本完成 DNA 序列分析的真核生物基因组比较

| 物 种 | 完成年份 | 总长度/Mb | 已完成总长的比例/% | 占常染色质的比例/% | 基因数/Mb |
|---|---|---|---|---|---|
| 酵母 | 1996 | 12 | 93 | 100 | 483 |
| 线虫 | 1998 | 96 | 99 | 100 | 197 |
| 果蝇 | 2000 | 116 | 64 | 97 | 117 |
| 拟南芥 | 2000 | 115 | 92 | 100 | 221 |
| 人类全基因组 Public Sequence | 2001 | 2 693 | 84 | 90 | 12 |
| 人类全基因组 Celera Sequence | 2001 | 2 654 | 83 | 93 ~ 99 | 15 |
| 小鼠基因组 | 2002 | 2 493 | 95 | 95 | 9 |
| 籼稻基因组 | 2002 | 466 | 92 | 98 | 110 |
| 大鼠基因组 | 2004 | 2 750 | 90 | 94 | 8 |
| 鸡基因组 | 2004 | 1 200 | 91 | 97 | 15 |
| 狗基因组 | 2005 | 2 410 | 99 | 100 | 8 |

不但基因组比较小，基因密度比较高，百万碱基对中含有 200 个或更多的基因(图 11–22)，而且异染色质的比例较低，基因组 90%以上由常染色质组成，而果蝇和人类基因组中异染色质的比例较高，占基因组的 20%~40%。研究果蝇全部四条染色体发现，其 Y 染色体几乎完全异染色质化，第四号染色体也大部分被异染色质化。果蝇 X 染色体在不同家系中变化最大，其异染色质化程度从 30%

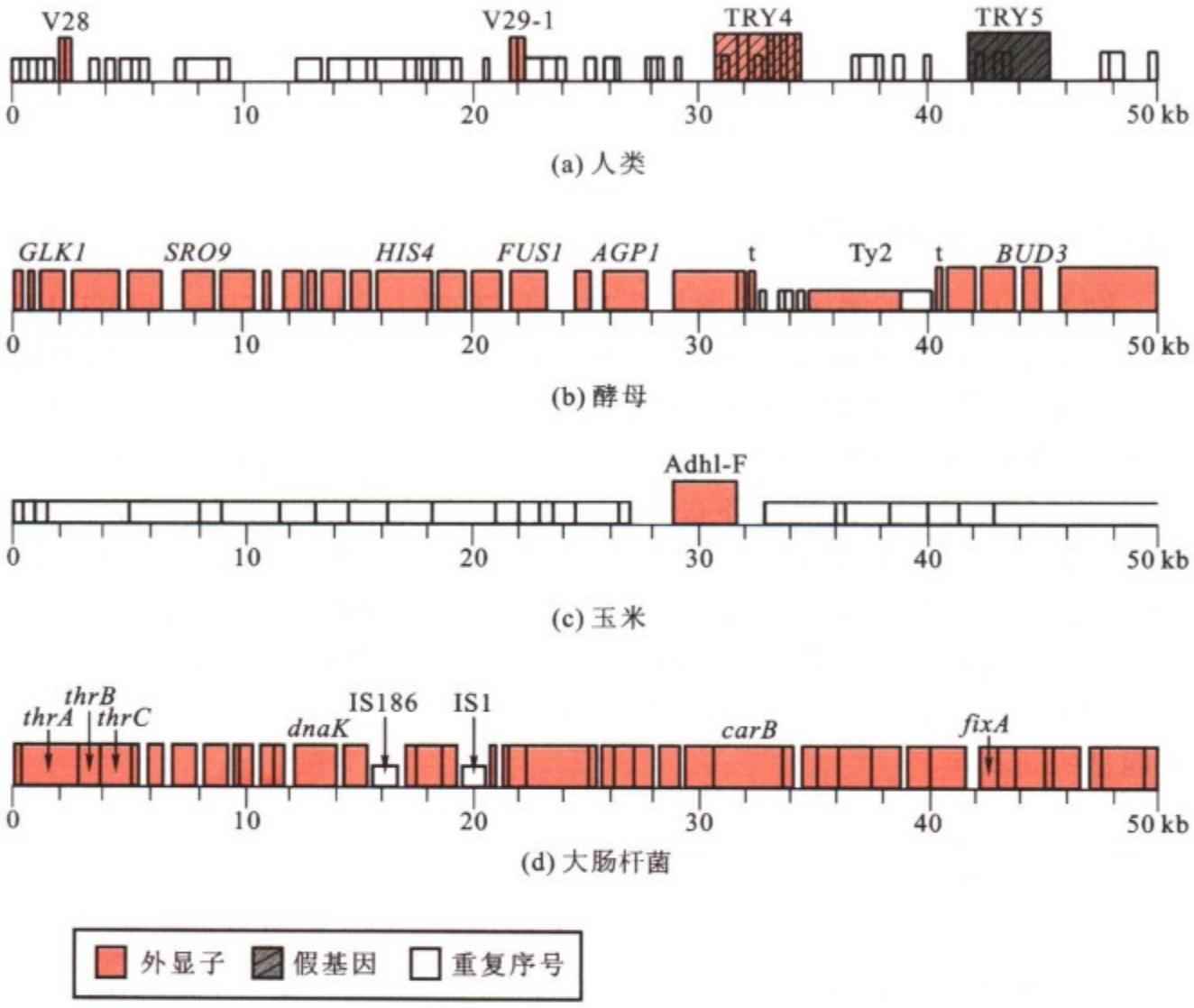



**图 11–22 部分典型真核和原核生物基因组成分分析** (a)人 β–T 细胞受体位点，在 50 kb 大片段中只有一个基因(*TRY4*，编码胰蛋白酶原)，一个假基因(*TRY5*)，两个基因片段(*V28*，*V29–1*)和 52 个存在于全基因组范围内的重复序列，编码功能基因的序列占总序列不到 3%。(b)酵母第Ⅳ号染色体中的一个片段，其中包括 26 个蛋白质编码基因，2 个 *tRNA* 基因，5 个存在于基因组范围内的重复序列，编码功能基因的序列占总序列的 66.4%，重复序列占 13.5%(在所有 16 条酵母染色体中，重复序列只有 3.4%)。在该 50 kb 序列中，所有基因都不带内含子，在整个酵母基因组中，一共只有 239 个内含子，而一个人类基因就可能有多达 100 个内含子。(c)在 50 kb 玉米基因组中，只有 1 个基因，乙醇脱氢酶 *I–F* 基因，其余几乎都是重复序列。在它的 5 000 Mb 基因组序列中，50%以上可能是重复序列。(d)大肠杆菌基因组中，50 kb 序列中可能有 43 个基因(占全序列的 85.9%)，许多基因之间甚至连一点空间都没有(*thrA* 和 *thrB* 之间只隔了一个碱基，*thrC* 基因直接位于 *thrB* 终止子的下游)。原核生物基因中没有内含子(少数古细菌基因中可能存在极少的内含子)。原核基因组中没有重复序列，但已发现存在某些插入序列(如 IS186 和 IS1)。在整个大肠杆菌 4 639 kb 序列中共发现不到 4 400 个编码基因。

到 50%左右不等(图 11-23)。

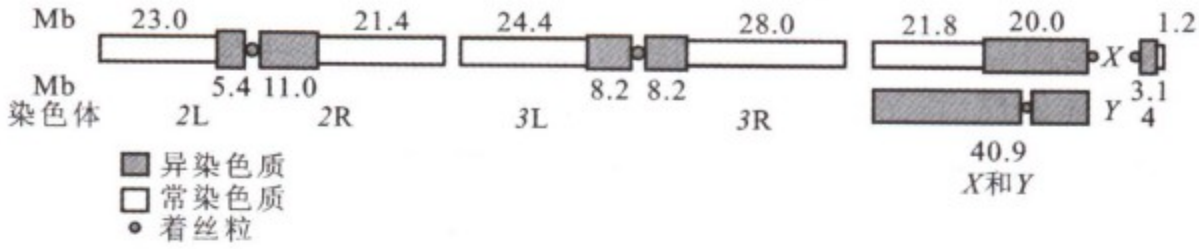



图 11-23 果蝇染色体中常染色质和异染色质含量比较

## 11.3.1 大肠杆菌基因组

K12 是第一个被测序的大肠杆菌基因组,有两个菌株(MG1655 和W3110)分别在 1996—1997 年间完成测序。1999—2001 年又先后完成了具有致病性的大肠杆菌 O157：H7 EDL933 和 SAKAI 菌株的基因组测序。表 11-5 是到目前为止已经完成测序的各种大肠杆菌基因组。大肠杆菌 K12：MG1655 全基因组是包含 4 639 221 bp 的双链环状分子，蛋白质编码基因占了整个基因组的 87.8%,*RNA* 基因占了 0.8%,0.7%为非编码的重复序列,其余约11%为调控或其他功能区。共含有 4 400 多个可能的编码基因,其中 3 700 个可以通过生物化学实验或者生物信息学分析得到合理的注释。另外大约 700 个基因中有 650 个与其他细菌基因组基因显示出较高的同源性，只有 50 个基因与其他任何物种没有明显的同源性。

表 11-5 到 2006 年为止已完成测序的大肠杆菌菌株

| 菌株名称 | 基因组大小/Mb | 基因数目 | 菌株名称 | 基因组大小/Mb | 基因数目 |
|---|---|---|---|---|---|
| O157_EDL933 | 5 528 | 5 349 | B7A | 5 202 | 4 637 |
| O157_Sakai | 5 498 | 5 361 | NMEC RS218 | 5 089 | ~4 900 |
| E22 | 5 516 | 4 788 | E2348 | 5 072 | 4 594 |
| E110019 | 5 384 | 4 746 | E24377A | 4 980 | 4 254 |
| B171 | 5 299 | 4 467 | UPEC 536 | ~4 900 | ~4 800 |
| 53638 | 5 289 | 4 783 | 101NA1 | 4 880 | 4 238 |
| 42 | 5 242 | 4 899 | HS | 4 643 | 3 689 |
| CFT073 | 5 231 | 5 379 | K-12_W3110 | 4 641 | 4 390 |
| H10407 | ~5 208 | ~5 000 | K-12_MG1655 | 4 639 | 4 254 |
| F11 | 5 206 | 4 467 | B03 | 4 629 | 4 387 |

研究发现,大肠杆菌基因组重复序列的排布可能与 DNA 的复制相关。如多拷贝重复排列的 *rRNA* 基因(*rrn*)位于环状染色体的中部,这个位置包含已知的 DNA 复制起始位点 oriC,其他许多与 rRNA 相关的 DNA 序列(TPIP)则位于环状染色体另一端的相应位置。此外,*rrn* 与 TPIP 序列对称分布在 oriC 两边的先

导链上。此外,*rrn* 操纵子的转录通常朝着远离复制起始位点的方向进行。

## 11.3.2 酵母基因组

酿酒酵母基因组是第一个测序完成的真核生物基因组,全基因序列于 1996 年完成。该基因组全长 13 040 000 bp,比单细胞的原核生物和古细菌大一个数量级。酿酒酵母是最小的真核基因组,基因密度为 1/2 kb。裂殖酵母比酿酒酵母基因密度稍大,其密度是 1/2.3 kb;而简单多细胞生物线虫的基因密度低达 1/30 kb。酵母基因组中共有 5 885 个蛋白质编码基因,另有大约 140 个 *rRNA*、40 个 *snRNA* (small nuclear RNA)和 275 个 *tRNA* 基因。酿酒酵母中约 4%的编码基因(大多为 *rRNA* 基因)有内含子,通常这些内含子位于靠近 *rRNA* 基因的起始部分。

比较不同酵母菌株发现,*S.kluyveri* 只有 7 条染色体,10 Mb;*S.castellii* 有 9 条染色体,9.72 Mb。*S.exiguus* 和 *S.cerevisiea* 的基因组最大,都包含 16 条染色体,总 DNA 量分别达到 18 355 000 和 13 030 000 bp,分别是已经测序的酵母菌株中第一和第二大的基因组(表 11-6)。

表 11-6 酵母各菌株染色体数目及单个染色体、基因组数据比较

| 菌株名称 | *S. dairenensis* | *S. castellii* | *S. servazzii* | *S. unisporus* | *S. exiguus* | *S. kluyveri* | *S. cerevisiea* |
|---|---|---|---|---|---|---|---|
| 染色体数目 | 9 | 9 | 12 | 12 | 16 | 7 | 16 |
| 单个染色体大小/kb | | | | | | | |
| 1 | 730 | 550 | 570 | 560 | 395 | 970 | 220 |
| 2 | 800 | 660 | 620 | 630 | 470 | 1 110 | 280 |
| 3 | 1 060 | 820 | 690 | 660 | 660 | 1 300 | 360 |
| 4 | 1 200 | 860 | 720 | 720 | 740 | 1 410 | 445 |
| 5 | 1 300 | 890 | 740 | 800 | 780 | 1 510 | 555 |
| 6 | 1 510 | 1 000 | 870 | 840 | 880 | 1 750 | 610 |
| 7 | 1 600 | 1 320 | 960 | 960 | 970 | 1 950 | 690 |
| 8 | 1 800 | 1 720 | 1 010 | 1 010 | 1 020 | — | 760 |
| 9 | 1 920 | 1 900 | 1 100 | 1 100 | 1 080 | — | 800 |
| 10 | — | — | 1 250 | 1 250 | 1 250 | — | 830 |
| 11 | — | — | 1 850 | 1 850 | 1 300 | — | 920 |
| 12 | — | — | 1 950 | 1 950 | 1 470 | — | 960 |
| 13 | — | — | — | — | 1 510 | — | 1 010 |
| 14 | — | — | — | — | 1 780 | — | 1 100 |
| 15 | — | — | — | — | 1 950 | — | 1 600 |
| 16 | — | — | — | — | 2 100 | — | 1 900 |
| 基因组大小/kb | 11 920 | 9 720 | 12 330 | 11 560 | 18 355 | 10 000 | 13 040 |

许多酵母染色体由交替的 GC 含量高和 GC 含量低的区段所组成,GC 含量的分布变化通常与这些染色体上的基因密度分布变化相一致。例如,在 3 号染

色体,碱基组成的周期性变化与染色体臂上的基因重组频率的变化相一致。GC富集区就是染色体臂中部的重组频率高的区域,AT富集区又与重组频率低的端粒和着丝粒区域序列对应。四个最小的染色体(1号,3号,6号和9号)表现出的平均重组率超过整个基因组平均水平的1.3~1.8倍。

### 11.3.3 拟南芥基因组

拟南芥基因组是世界上第一个被测序的植物基因组,也是继线虫、果蝇之后第三个被测序的多细胞生物。拟南芥基因组全长约为115.4 Mb,其中包含部分着丝粒区域的序列。基因组分析表明,在整个拟南芥进化过程中可能包括一次全基因组的复制以及广泛的基因缺失和重复复制,基因组中含有许多从蓝藻中迁移过来的DNA片段。2000年发表基因组数据时宣布,该基因组可能共有25 498个编码基因。通过最近6年多的努力,拟南芥5条染色体共编码约30 000个基因(表11-7)。据估计,这个数字可能不会增加太多了。与酵母线虫、果蝇相比,拟南芥基因组中编码转录调控因子的基因数量显著增加(表11-8),从一个侧面反映了高等植物的复杂性。

表11-7 TAIR与NCBI所公布的拟南芥各染色体基因数目比较

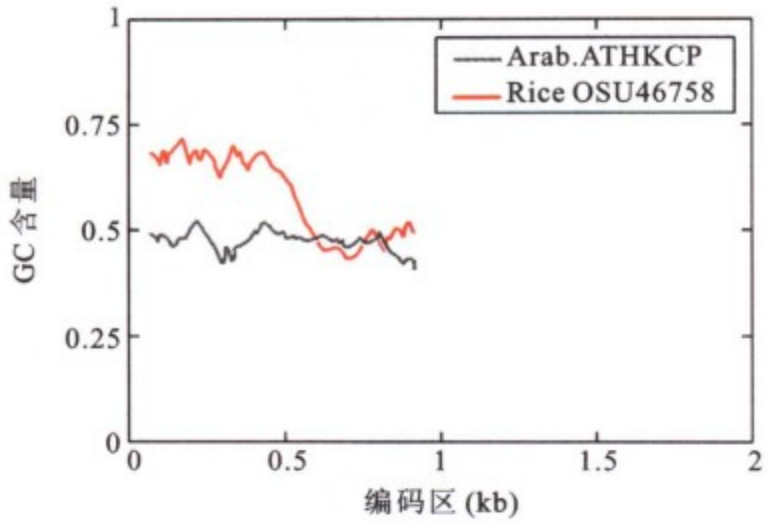
| | TAIR | NCBI | | TAIR | NCBI |
|---|---|---|---|---|---|
| 1号染色体 | 7 437 | 8 018 | 4号染色体 | 4 574 | 4 848 |
| 2号染色体 | 4 753 | 5 152 | 5号染色体 | 6 769 | 7 120 |
| 3号染色体 | 5 825 | 6 362 | 合计 | 29 388 | 31 500 |

表11-8 主要真核生物中转录调控因子数量比较

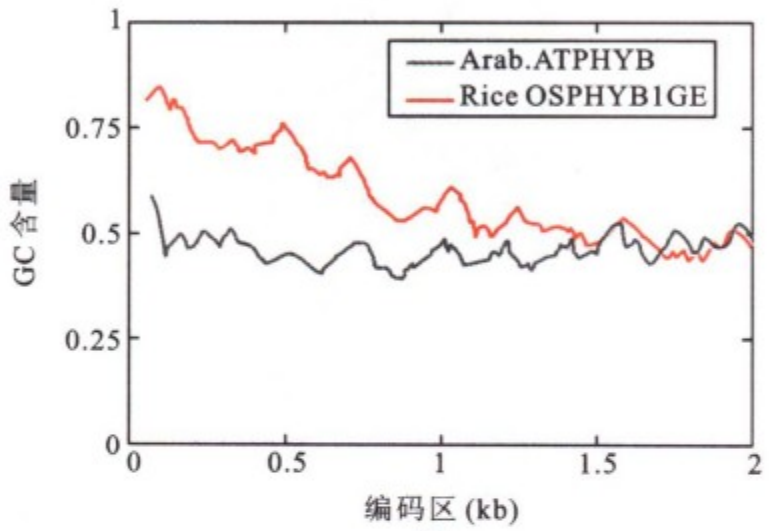
| 物种名 | 基因数目 | 转录因子数量 | 比例 |
|---|---|---|---|
| 拟南芥 | 约29 388 | 1 533 | 5.9 |
| 酵母 | 约5 885 | 209 | 3.5 |
| 线虫 | 约18 891 | 669 | 3.5 |
| 果蝇 | 约13 379 | 635 | 4.5 |

### 11.3.4 水稻基因组

籼稻和粳稻基因组几乎在2002年被同时完成。水稻共有12条染色体,籼稻基因组(常染色体部分)为466 Mb,粳稻为420 Mb,是拟南芥基因组的3.7倍,仅为人类基因组的15%。研究发现几乎所有的水稻基因至少有一个外显子区域表现出较高的GC含量,而在同一个基因内部5′区的GC含量要比3′区高出25%(图11-24)。另外水稻基因的平均大小只有4 500 bp,远小于人类基因的平均大小(27 kb)。水稻cDNA中转座子的含量约为1%,也低于人类的4%,表明了水稻在进化中的稳定性。基因组水平的同源性分析表明,80%的拟南芥基因可以在水稻中找到同源基因,其同源部分也占到蛋白质长度的80%,其中60%的氨基酸序列完全一样。相反,只有50%的水稻基因可以在拟南芥中找到




(a)




(b)

**图 11-24 水稻与拟南芥基因内部 GC 含量变化比较** (a)拟南芥 $K^+$通道蛋白(HKCP)与水稻同源基因 *OSU46758* 内部 GC 含量变化比较;(b)拟南芥光敏色素 B 基因(*PHYB*)与水稻同源基因 *OSPHYB1GE* 内部GC 含量变化比较。图示无论是 1 kb 大小的 *HKCP* 基因还是长度超过 10 kb 的*PHYB* 基因，水稻基因 5′的 GC 含量均要明显高于拟南芥基因。这两种植物基因编码区 3′的 GC 含量则大致相当。

同源基因,说明在高等单子叶植物进化过程中可能获得了大量其他来源的基因。

### 11.3.5 家蚕基因组

家蚕是一种独特的完全依赖于人类存活繁衍的驯养昆虫。用于家蚕基因组测序的 DNA 以及构建相应 cDNA 文库的 mRNA 来自一个纯驯养品种——大造(图 11-25)。因为容易在实验室饲养、有可利用的来自基因同源近等系的突变和生物学上的大量信息,家蚕一直是昆虫遗传学上的模式生物之一,其影响可能仅次于果蝇。家蚕有 400 个可见的表现型,包括许多重要的生物学特性如丝产量、饮食特性、抗病能力、胚胎形成和性征变异等,其中大约有 200 个分布在已知的连锁群。家蚕一共有 28 条染色体,测序的基因组大小为 428.7 Mb,是果蝇基因组的 3.6 倍,是蚊子的 1.54 倍。注释的家蚕基因为 18 510 个,大于果蝇的 13 379 的基

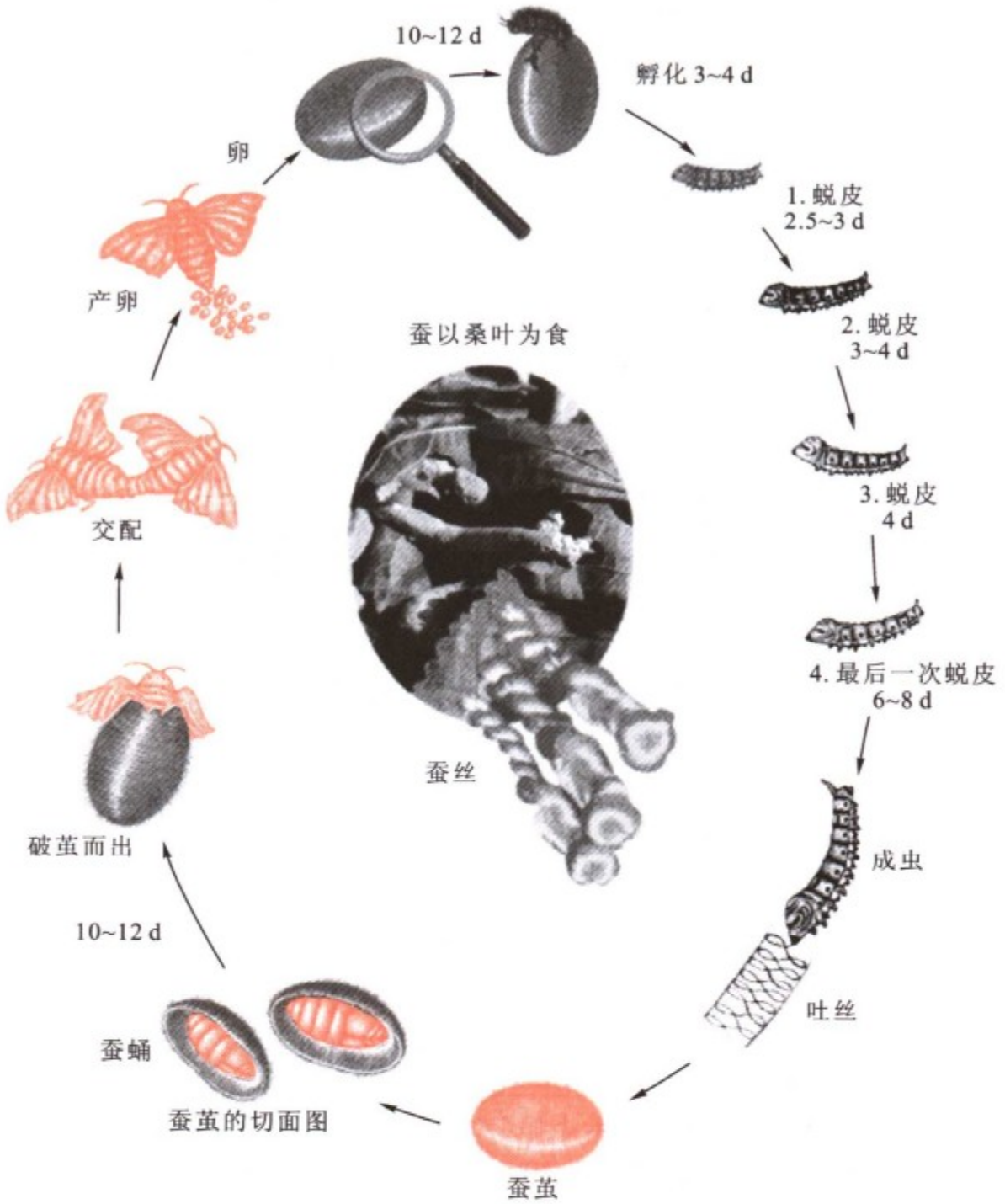



图 11-25 家蚕(*Bombyx mori*)的生命周期

因数。每个家蚕基因平均带有 1.15 个外显子,而果蝇基因几乎没有断裂。而且由于转座子主要插入内含子部分,家蚕的基因显著大于相应的果蝇基因(图 11-26)。

### 11.3.6 鸡基因组

科学家选择现代家鸡的祖宗野生红原鸡(*Gallus gallus*)作为基因组框架图的原型。鸡基因组草图由全基因组鸟枪法完成测定,基因组大小只有人类的 1/3 左右,但鸡基因组 DNA 的重复度极低,只有 11%(一般哺乳动物都高达 40%~50%)。除了一对性染色体外,鸡有 38 对常染色体,其中 5 个大染色体,5 个中染色体,28 个小染色体(图 11-27)。因为每个染色体臂至少有一个交叉(crossover),所以,小染色体基因重组的概率较高。研究发现,大染色体交叉率为 2.8 cM/Mb,小染色体交叉率为 6.4 cM/Mb,而大多数人类染色体的交叉率仅为

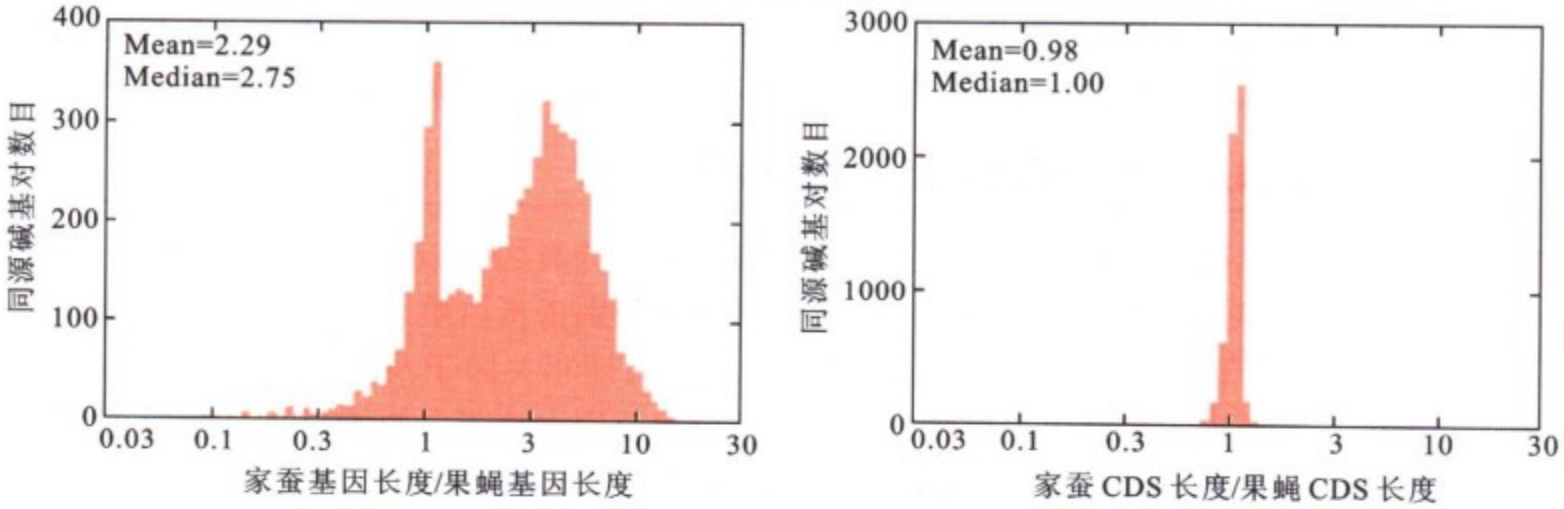



图 11-26 家蚕与果蝇同源基因大小的比较 (a) 5 879 个家蚕-果蝇同源基因去除内含子前 DNA 序列长度比较;(b)去除内含子后编码区(CDS)长度比较。

图 11-27 家鸡(*Gallus gallus domesticus*)染色体结构图

1~2 cM/Mb,这就使鸡成为研究遗传连锁的理想对象。另外,鸡基因组比其他哺乳动物基因组要紧凑得多,它只有 10 亿个 DNA 碱基对却编码了 2~2.3 万个基因,而人类基因组含有三十亿个碱基对却只编码了约 3 万个基因。鸡的基因数量与哺乳动物相当,但它的假基因数量很少,只有 75 个,而哺乳动物一般有 15 000 个左右。该现象与大量存在的重复元素(LINE)CR1 反转录酶的高度特异性密切相关,并使鸡种间和种内的变异程度相当接近。

## 11.4 比较基因组学及相关研究

基因组的序列主要可被分为三类:(一) 通过比较确知其生理功能的;(二)在数据库中有相匹配的蛋白质序列,但并不知道其功能的;(三)在现有数据库中找不到任何相匹配的蛋白质序列的新基因。由于生物在进化上是相互关联的,对一种生物的研究可以为其他生物提供有价值的信息。比较基因组学(comparative genomics)的威力在于它能根据对一种生物相关基因的认识来理解、诠释甚至克隆分离另一种生物的基因。远缘基因组间的比较为认识生物学机制的普遍性,寻找研究复杂生理和病理过程所需的实验模型提供了理论依据,而近缘基因组间的比

较则为认识基因结构与功能等细节提供了参数。因此,为充分理解人类基因组,必须对一系列近缘和远缘的模式生物进行基因组程度上的比较分析工作。

### 11.4.1 基因组数据的挖掘与分析

研究大肠杆菌 *K-12* 基因组发现,除了尚有 38%以上的未知蛋白质之外,与物质运转和能量代谢相关的蛋白质含量分别占蛋白质总量的 9%左右,各种功能性酶、细胞结构蛋白、调控蛋白、细胞周期相关因子及参与蛋白质合成、参与重要中间物合成与代谢等过程的蛋白质分别占总蛋白的 4%以上,参与氨基酸合成及代谢,参与 DNA 合成及代谢的蛋白质也都达到总蛋白的 3%左右(表 11-9)。

表 11-9 大肠杆菌所编码的蛋白质分类

| 功 能 | 数量 | 占蛋白质总量的比例/% |
|---|---|---|
| 调控蛋白 | 178 | 4.15 |
| 细胞结构蛋白 | 224 | 5.22 |
| 膜蛋白 | 13 | 0.3 |
| 外源蛋白 | 87 | 2.03 |
| 物质运转相关蛋白 | 427 | 9.95 |
| 能量代谢相关蛋白 | 373 | 8.70 |
| DNA 合成与代谢 | 115 | 2.68 |
| 转录及 RNA 合成、代谢与修饰 | 55 | 1.28 |
| 翻译及蛋白质修饰 | 182 | 4.24 |
| 细胞周期相关因子 | 188 | 4.38 |
| 辅基、辅因子及其载体 | 103 | 2.40 |
| 伴侣蛋白 | 9 | 0.21 |
| 核苷酸的合成与代谢 | 58 | 1.35 |
| 氨基酸的合成与代谢 | 131 | 3.06 |
| 脂肪酸及磷脂的合成与代谢 | 48 | 1.12 |
| 重要中间物合成与代谢 | 188 | 4.38 |
| 酶 | 251 | 5.85 |
| 其他(功能已知蛋白) | 26 | 0.61 |
| 未知蛋白 | 1 632 | 38.06 |

人类基因组研究还发现,人类基因的平均长度为 27 kb 左右,含有 8.8 个长约 145 bp 的外显子,内含子的长度大大超过外显子,达到 3 365 bp 左右。人类基因的 3′非翻译区(UTR)的平均长度为 770 bp,其 5′非翻译区的平均长度为 300 bp,可译框架的平均长度只有 1 340 bp,编码 447 个氨基酸(表 11-10)。

对原始的流感嗜血杆菌基因组与酵母、线虫、果蝇和拟南芥进行比较以后发现,原始生物细胞中单拷贝基因较多,在流感嗜血杆菌中单拷贝基因占 88.8%,在酵母中占 71.4%,在果蝇中占 72.5%,在线虫中占 55.2%,而在高等植物拟南芥中只占约 35.0%(表 11-11)。

表 11-10 人类基因组数据库中蛋白质编码基因的部分重要参数比较

| 性　质 | 平均值 | 性　质 | 平均值 |
|---|---|---|---|
| 外显子长度/bp | 145 | 5′非翻译区 | 300 |
| 外显子个数 | 8.8 | 可译框架的长度/bp | 1 340 |
| 内含子长度/bp | 3 365 | 核基因长度/kb | 27 |
| 3′非翻译区 | 770 | | |

表 11-11 不同物种中单拷贝基因的数量及占基因总数百分比

| 物　种 | 单拷贝基因个数 | 单拷贝基因占基因总量的% |
|---|---|---|
| 流感嗜血杆菌 | 1 587 | 88.8 |
| 酵母 | 5 105 | 71.4 |
| 果蝇 | 10 736 | 72.5 |
| 线虫 | 14 177 | 55.2 |
| 拟南芥 | 11 601 | 35.0 |

## 11.4.2 基因组数据的比较研究

尿殖道支原体是已知最小的基因组,可依此确定能自我复制的细胞必需的一套最少的核心基因。流感嗜血杆菌的基因组为 1.83 Mb,而尿殖道支原体的基因组只有 0.58 Mb,两者相差 3 倍多,那么,基因组大小影响了基因数目还是基因尺度?流感嗜血杆菌基因大小平均 900 bp,尿殖道支原体的基因为 1 040 bp,基因大小差不多。流感嗜血杆菌中平均 1 042 bp 有 1 个基因,尿殖道支原体中平均 1 235 bp 有 1 个基因。可见基因组尺度减小并不引起基因密度的增加和基因本身尺寸的减小。两者的差别在于基因数量上,流感嗜血杆菌基因组有 1 743 个 ORF,而尿殖道支原体只有 470 个 ORF(表 11-12)。通过对尿殖道支原体与流感嗜血杆菌这两个亲缘关系较远的生物基因组的比较,选取其共同的基因(共 240 个),再加上一些其他基因,最后组成一套含256 个基因的最小基因组。

表 11-12 流感嗜血杆菌和尿殖道支原体各类主要基因比较

| 分　类 | 流感嗜血杆菌 | 尿殖道支原体 | 分　类 | 流感嗜血杆菌 | 尿殖道支原体 |
|---|---|---|---|---|---|
| 总 ORF 数 | 1 727 | 470 | 脂肪代谢 | 25 | 6 |
| 氨基酸合成 | 68 | 1 | DNA 复制、重组和修复 | 87 | 42 |
| 辅基等的合成 | 54 | 5 | 蛋白质高级结构形成 | 6 | 7 |
| 核苷酸合成 | 53 | 29 | 调控蛋白 | 64 | 7 |
| 细胞膜合成与装配 | 84 | 27 | 转录 | 27 | 22 |
| 能量代谢 | 112 | 31 | 翻译 | 141 | 101 |
| 糖代谢 | 30 | 6 | 吸收与转运 | 123 | 34 |

通过流感嗜血杆菌能量代谢类群的 ORF 分析，了解到在这种生物中缺乏三羧酸循环(TCA)中必需的三个酶，即柠檬酸合成酶基因、异柠檬酸脱氢酶基因和顺乌头酸酶基因。由此推断流感嗜血杆菌 TCA 缺失，不能合成谷氨酸，因为谷氨酸的供体是 TCA 的中间产生物 α-酮戊二酸。

古细菌产甲烷球菌与原核生物有着共同的染色体组织与结构，如环状基因组、基因的操纵子结构等，其能量产生和固氮基因与原核生物也有很高的同源性。该基因组中与细胞分裂有关的蛋白质及 20 多个编码无机离子运输蛋白的 ORF 与细菌基因同源，而且其调控模式也类似于原核生物。然而，产甲烷球菌在细胞遗传信息传递，尤其是转录和翻译系统，以及分泌系统方面与真核生物同源，说明该细菌与真核生物亲缘关系较近。比较基因组学提供的结果表明，在进化系统树上，古细菌与真核生物亲缘关系比原核生物更近。在自养生物的三个分支，细菌、古细菌和真核生物中，细菌的分化发生较早。

## 思考题

1. 简述人类基因组计划的科学意义。
2. 请分析人类基因组计划的发展趋势。
3. 试述大肠杆菌基因组和真核生物基因组的主要区别。
4. 列出 LINEs 与 SINEs 的异同。
5. 区分遗传图谱和物理图谱，试述这两项技术的优缺点。
6. 什么是 FISH 技术？如何利用它来构建基因组的物理图谱？
7. 鸟枪测序法的技术原理是什么？
8. 什么是蛋白质组学？什么是转录组学？简述这两个领域里最主要的研究手段。

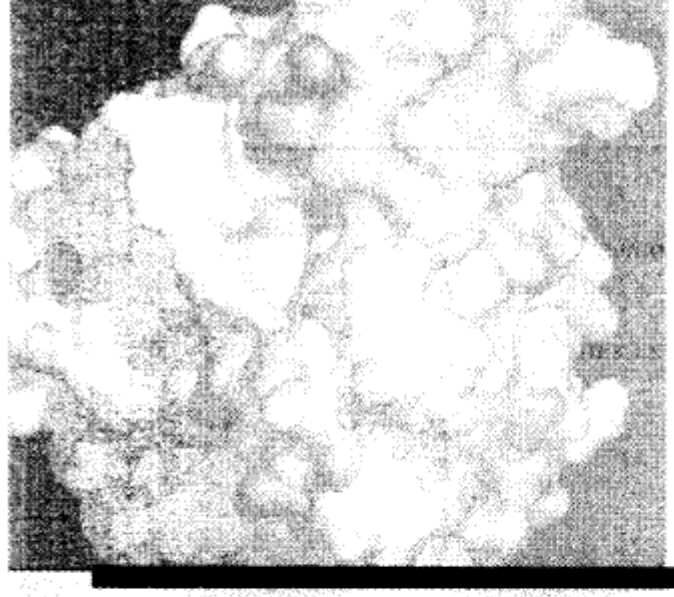
## 主要参考文献


1. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature, 2000, 408: 796–815.
2. Blattner F R, Plunkett G 3rd, Bloch C A, *et al*. The complete genome sequence of Escherichia coli K-12. Science, 1997, 277: 1453–1474.
3. Goffeau A, Barrell B G, Bussey H et al. Life with 6000 genes. Science, 1996, 274: 563–567.
4. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides N C. The Genomes On Line Database (GOLD)v.2: a monitor of genome projects worldwide. Nucleic Acids Res, 2006, 34: 332–334.
5. Sachidanandam R, Weissman D, Schmidt S C, *et al*. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature, 2001, 409: 928–933.
6. Venter J C, Adams M D, Sutton G G, *et al*. Shotgun sequencing of the human

genome. Science, 1998, 280: 1540–1542.

7. Xia Q, Zhou Z, Lu C, *et al.* A draft sequence for the genome of the domesticated silkworm (Bombyx mori). Science, 2004, 306: 1937–1940.

8. Yu J, Hu S, Wang J, *et al.*, A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science, 2002 296: 79–92.



# 名词解释

**ABC 模型**:即控制花形态发生的模型。该模型把四轮花器官同时发生作为基本前提,强调花形态突变体产生不同花器官的生理位置变化。该模型中正常花的四轮结构的形成是由三组基因 *A*、*B*、*C* 共同作用完成的,每一轮花器官特征的决定分别依赖于 *A*、*B*、*C* 三组基因中的一组或两组基因的正常表达。*A* 组基因控制萼片、花瓣的发育,*B* 组基因控制花瓣、雄蕊的发育,*C* 组基因控制雄蕊、心皮的发育。*A*、*C* 组基因互相拮抗,抑制对方在自身所控制的区域中表达,如其中任何一组或更多的基因发生突变而丧失功能,花的形态就出现异常。

**AP 位点(AP site)**:所有细胞中都带有不同类型、能识别受损核酸位点的糖苷水解酶,它能特异性切除受损核苷酸上的 *N*-*β* 糖苷键,在 DNA 链上形成去嘌呤或去嘧啶位点,统称为 AP 位点。

**cDNA(complementary DNA)**:在体外以 mRNA 为模板,利用反转录酶和 DNA 聚合酶合成的一段双链 DNA。

**C 值(C value)**:通常是指一种生物单倍体基因组 DNA 的总量,以每细胞内的皮克(pg)数表示。

**C 值反常现象(C value paradox)**:也称 C 值谬误。指 C 值往往与种系的进化复杂性不一致的现象,即基因组大小与遗传复杂性之间没有必然的联系,某些较低等的生物 C 值却很大,如一些两栖动物的 C 值甚至比哺乳动物还大。

**Dane 颗粒**:HBV 完整颗粒的直径为 42 nm,称为 Dane 颗粒,由外膜和核壳组成,有很强的感染性。

**DNA(deoxyribonucleic acid)**:脱氧核糖核酸,是世界上所有已知高等真核生物和绝大部分低等生物的遗传物质。

**DNA 的半保留复制(semi-conservative replication)**:DNA 在复制过程中,每条链分别作为模板合成新链,产生互补的两条链。这样新形成的两个 DNA 分子与原来 DNA 分子的碱基顺序完全一样。因此,每个子代分子的一条链来自亲代 DNA,另一条链则是新合成的,这种复制方式被称为 DNA 的半保留复制。

**DNA 的半不连续复制(semi-discontinuous replication)**:DNA 复制过程中前导链的复制是连续的,而另一条链,即后随链的复制是中断的、不连续的。

**DNA 甲基化**:CpG 二核苷酸(CpG 岛)通常成串出现在 DNA 上,在甲基转移酶的作用下,胞嘧啶(C)的第 5 位碳原子能被修饰加上甲基。DNA 甲基化除形成 5-甲基胞嘧啶(5-mC)之外,还能产生少量的 $N^6$-甲基腺嘌呤($N^6$-mA)及 7-甲基鸟嘌呤。

**DNA 聚合酶 (DNA polymerase)**: 一种催化由脱氧核糖核苷三磷酸合成 DNA 的酶。因为它以 DNA 为模板,所以又被称为依赖于 DNA 的 DNA 聚合酶。不同种类的 DNA 聚合酶可能参与 DNA 的复制和/或修复。

**DNA 酶 I 超敏感位点**: 染色质中特殊的一段长约 200 bp、甲基化程度较低且对 DNase Ⅰ高度敏感的 DNA 序列,一般在转录起始点附近或者相关部位。

**DNA 拓扑异构酶(DNA topoisomerase)**:能在闭环 DNA 分子中改变两条链的环绕次数的酶,它的作用机制是首先切断 DNA,让 DNA 绕过断裂点以后再封闭形成双螺旋或超螺旋 DNA。

**DNA 重组技术(recombinant DNA technology)**:又称基因工程(genetic engineering),将不同的 DNA 片段(如某个基因或基因的一部分)按照预先的设计定向连接起来,在特定的受体细胞中与载体同时复制并得到表达,产生影响受体细胞的新的遗传性状的技术。

**GU-AG 法则(GU-AG rule)**:多数细胞核 mRNA 前体中内含子的 5′边界序列为 GU,3′边界序列为 AG。因此,GU 表示供体衔接点的 5′端,AG 代表接纳体衔接点的 3′端序列。习惯上,把这种保守序列模式称为 GU-AG 法则。

**HLA**:即人类白细胞抗原,是受遗传控制的个体特异性抗原,广泛分布于皮肤、肾、脾、肺、肠和心等组织器官有核细胞的细胞膜上,是到目前为止免疫反应中最复杂、最具多态性的体系。

**$Ig^{\circ}$,$Ig^{+}$,$Ig^{-}$**:为区分免疫球蛋白基因的各种状态,科学上将未发生重排的种系构型等位基因称为 $Ig^{\circ}$,将发生重排并具备功能表达的等位基因称为 $Ig^{+}$,将发生无效重排的无活性等位基因称为 $Ig^{-}$。

**MADS-box**:在金鱼草、拟南芥等多种植物中存在一个参与花器官发育和分化的基因家族的保守区域,该家族的成员可能参与不同的分化过程。根据这几个家族基因名称(*MCM1*、*AG*、*DEFA* 和 *SRF*)的第一个字母,将该保守区命名为 MADS-box。

**N 区**:Ig 重链基因重排过程中, 由于 DNA 聚合酶或脱氧核酸转移酶的作用,有时会在结合处插入一个脱氧核苷酸,形成所谓“N 区”插入序列,致使重链

V 区重排后形成 $V_HND_HNJ_H$ 基因单位。上述重排机制使重链 V 区变得更为复杂。

**P 转座子(P-element)**:是一种能够诱发杂种不育(hybrid dysgenesis)的果蝇转座子，果蝇中几乎所有的杂种不育都是由于 P 转座子插入基因组 W 位点而引起的。所有 P 转座子两翼都有 31 个碱基的倒置重复序列,P 转座子转座导致靶 DNA 复制产生 8 碱基正向重复序列。

**RACE(rapid amplification of cDNA ends,cDNA 末端的快速扩增)**:是利用 PCR 技术在已知部分 cDNA 序列的基础上特异性克隆其 5′端或 3′端缺失序列的方法。

**RAPD(random amplified polymorphic DNA,随机扩增的多态性 DNA)**:用长度为 10 或 11 个碱基的单一固定序列作引物扩增基因组 DNA,随机获得大小不同的 DNA 片段。因其具有种质特异性而常用作遗传学的分子标记。

**RFLP (restriction fragment length polymorphism，限制性片段长度多态性)**:使用限制性 DNA 内切酶消化某个 DNA 分子后出现的长度多样性。因其具有种质特异性而常用作遗传学的分子标记。

**RNA 的编辑(RNA editing)**:是某些 RNA,特别是 mRNA 前体的一种加工方式,如插入、删除或取代一些核苷酸残基,导致 DNA 所编码的遗传信息发生改变,因为经过编辑的 mRNA 序列发生了不同于模板 DNA 的变化。

**RNA 的再编码(RNA recoding)**:是指 RNA 编码和读码方式的改变。

**RNA 干涉(RNA interference,RNAi)**:是利用双链小 RNA 高效、特异性降解细胞内同源 mRNA,从而阻断体内靶基因表达,使细胞出现靶基因缺失表型的方法。

**RNA 的剪接(RNA splicing)**:从 mRNA 前体分子中切除被称为内含子(intron)的非编码区,并使基因中被称为外显子(exon)的编码区拼接形成成熟 mRNA 的过程就称为 RNA 的剪接。

**RNA 聚合酶(RNA polymerase)**:使用 DNA 作为模板合成 RNA 的酶,也称为 DNA 依赖性 RNA 聚合酶。

**SARS-CoV**:一种新型冠状病毒,属于冠状病毒科冠状病毒属,是一种大的、有包膜的正链 RNA 病毒,其基因组由约 30 000 个核苷酸组成,直径约 70~140 nm,电镜下可见电子密度致密的核心和围绕的包膜,包膜上有冠状的刺突。对脂溶剂敏感,戊二醛、甲醛、过氧化氢、表面活性剂、紫外线照射等可使病毒失去感染力。可导致传染性非典型肺炎(SARS,severe acute respiratory syndrome)——严重急性呼吸系统综合症。

**SD 序列(Shine-Dalgarno sequence)**:存在于原核生物起始密码子 AUG 上游 7~12 个核苷酸处的一种 4~7 个核苷酸的保守片段,它与 16S rRNA 3′端反向互补，所以可将 mRNA 的 AUG 起始密码子置于核糖体的适当位置以便起始翻译作用。根据首次识别其功能意义的科学家命名。

**SNP(single nucleotide polymorphism)**:单核苷酸多态性,指基因组 DNA 序列中由于单个核苷酸(A,T,C 和 G)的突变而引起的物种多态性。

**T-DNA**:根癌农杆菌 Ti(tumor inducing)或 Ri(root inducing)质粒中的一段

DNA 序列,可以从农杆菌中转移并稳定整合到植物基因组中,是植物分子生物学中广泛应用的遗传转化载体。

**ρ 因子**:是一个相对分子质量为 $2.0\times10^5$ 的六聚体蛋白,它能水解各种核苷三磷酸,是一种 NTP 酶,它通过催化 NTP 的水解促使新生 RNA 链从三元转录复合物中解离出来,从而终止转录。

**σ 因子(sigma factor)**:是原核生物 RNA 聚合酶全酶的一个亚基,是聚合酶的别构效应物,帮助聚合酶专一性识别并结合模板链上的启动子,起始基因转录。

## A

**癌(cancer)**:是一种无限制向外周扩散、浸润现象。癌症患者的主要特征是发病组织或器官的细胞生长分裂失控,并由原发部位向其他部位播散。如不能控制这种细胞播散,将侵犯要害器官并引起衰竭,最终导致有机体死亡。

**安慰性诱导物(gratuitous inducer)**:指的是与转录调控中实际诱导物相似的一类高效诱导物,但不是该诱导酶的底物。

**氨酰-tRNA 合成酶(aminoacyl-tRNA synthetase)**:是一类催化氨基酸与 tRNA 相结合的特异性酶。

## B

**摆动假说(wobble hypothesis)**:Crick 为解释反密码子中某些稀有成分(如 I)的配对以及许多氨基酸有 2 个以上密码子的问题而提出的假说。

**比较基因组学(comparative genomics)**:在基因组图谱和序列分析的基础上,对已知基因和基因组结构进行比较,了解基因的功能、表达调控机制和物种进化过程的学科。

**编码链(coding strand)**:指 DNA 双链中与 mRNA 序列(除 T/U 替换外)和方向相同的那条 DNA 链,又称有意义链(sense strand)。

**表达序列标签(expressed sequence tag,EST)**:通过对随机选择的 cDNA 克隆进行 5′或 3′端一次测序所获得的核苷酸序列,代表了完整基因的一小部分。EST 一般来源于用特定环境下某个组织总 mRNA 所构建的 cDNA 文库,因此 EST 数量也能在一定程度上说明该组织中各基因的表达水平。

**病毒癌基因(v-onc)**:一些病毒能引发癌症,原因是病毒基因片段进入被感染者基因组中,导致被感染者自身基因行为发生改变而出现细胞癌变。科学上,将这种可以引发癌症的病毒基因片段称为“病毒癌基因”。

**病毒载体**:对病毒基因组进行改造后获得的适合用于基因治疗的载体,主要是删除病毒的致病基因,使其对机体没有致病力,该载体应能携带外源基因、装配成病毒颗粒并介导外源基因的转移和表达。

## C

**操纵子(operon)**:是指原核生物中由一个或多个相关基因以及转录翻译调

控元件组成的基因表达单元。

**插入序列(insertional sequence,IS)**:是最简单的转座子,是细菌的一小段可转座元件,它不含有任何宿主基因而常被称为插入序列,它们是细菌染色体或质粒 DNA 的正常组成部分。

**长末端重复序列(LTR)**:指反转录病毒基因组两端的长末端重复序列(long terminal repeats),不编码蛋白质,但含有启动子、增强子等调控元件,病毒基因组的 LTR 整合到细胞原癌基因邻近处,使这些原癌基因在 LTR 强启动子和增强子的作用下被激活,将正常细胞转化为肉瘤细胞。

**超螺旋(superhelix,supercoil)**:双螺旋 DNA 进一步扭曲盘绕所形成的特定空间结构,是 DNA 高级结构的主要形式,可分为正超螺旋与负超螺旋两大类。按 DNA 双螺旋的相反方向缠绕而成的超螺旋称为负超螺旋,反之,则称为正超螺旋。所有天然的超螺旋 DNA 均为负超螺旋。

**成对规则基因**:参与果蝇体节精细结构形成的基因,一般以两个体节为单位相互间隔一个副体节表达,其分布具有周期性。功能是把间隙基因确定的区域进一步分化成体节。成对规则基因的表达是胚胎出现分节的最早标志之一,沿前-后轴形成一系列斑马纹状的表达条带,把胚胎分为预定的体节。

**重叠基因(overlapping gene,nested gene)**:具有部分共用核苷酸序列的基因,即同一段 DNA 携带了两种或两种以上不同蛋白质的编码信息。重叠的部分可以在调控区,也可以在结构基因区。常见于病毒和噬菌体基因组中。

**重组种系理论**:是有关生物体为何能产生上百万种 Ig 分子的假说。重组种系理论认为,V 区和 C 区不同片段在 DNA 分子水平上的各种排列组合是形成 Ig 分子多态的根本原因。

**错配修复(mismatch repair)**:是对 DNA 错配区的修复。通过母链甲基化原则找出区分母链与子链从而修正子链上错配的碱基。

**错义突变(missense mutation)**:由于结构基因中某个核苷酸的变化使一种氨基酸的密码变成另一种氨基酸的密码。

## D

**代谢物阻遏效应(catabolite repression)**:有葡萄糖存在时,不论诱导物存在与否,操纵子都没有转录活性,结构基因都不表达。

**单倍型(haplotype)**:是指位于染色体上某一区域的一组相关联的 SNP 等位位点。

**单核苷酸的多态性(single nucleotide polymorphism,SNPs)**:单核甘酸多态性指在基因组中不同个体的 DNA 序列上的单个碱基差异。同一位置上的每个碱基类型叫做一个等位位点。

**单顺反子 mRNA(monocistronic mRNA)**:只编码一个蛋白质的 mRNA 称为单顺反子 mRNA。

**蛋白激酶**:催化 ATP 的 γ-磷酸基团转移到蛋白质分子中的酶。

**蛋白质磷酸化**:指由蛋白质激酶催化的把 ATP 或 GTP 上 γ 位的磷酸基转

移到底物蛋白质氨基酸残基上的过程，是生物体内一种普遍的调节方式，在细胞信号转导的过程中起重要作用。

**蛋白质乙酰化**：在乙酰基转移酶的作用下，在蛋白质特定的位置添加乙酰基的过程，是细胞控制基因表达、蛋白质活性或生理过程的一种机制。

**蛋白质组(proteome)与蛋白质组学(proteomics)**：蛋白质组是指一个基因组所表达的全部蛋白质，而蛋白质组学则是指在蛋白质组水平上研究蛋白质的特征，包括蛋白质的表达水平、翻译与修饰、蛋白与蛋白相互作用等，并由此获得关于疾病发生、发展及细胞代谢等过程的整体认识。

**等位基因(allele)**：指位于一对同源染色体的相同位置上控制某一性状的不同拷贝。不同的等位基因产生具有遗传特征的变化，例如，发色或血型等。等位基因控制相对性状的显隐性关系及遗传效应。

**等位排斥与同型排斥**：淋巴细胞中产生免疫球蛋白的基因位于两条同源染色体上，而免疫球蛋白基因的表达只发生在其中一条上，这种因为一条染色体上的基因表达而抑制另一条染色体上相同基因表达的现象称为等位基因排斥。而同型排斥则是指 B 淋巴细胞的轻链表达时，只生成一种链(κ 链或是 λ 链)，不会同时表达 κ 链和 λ 链。

**等位位点**：染色体 DNA 同一位置上的每个碱基类型叫做一个等位位点。

**多聚 A 位点(poly A 位点)**：在多聚 A 聚合酶的催化下，在 mRNA 前体 3′端接上一个约200~250 个腺嘌呤核苷酸(A)的尾巴，起到稳定 mRNA 的作用。

**多顺反子 mRNA(polycistronic messenger RNA)**：一种能作为两种或多种多肽链翻译模板的信使 RNA，由 DNA 链上的邻近顺反子所界定。

## E

**二组分系统(two-component system)**：由位于细胞质膜上的传感蛋白(该蛋白常常具有激酶活性)以及位于细胞质中的应答调节蛋白组成。传感激酶常在受到膜外环境信号刺激时被磷酸化，并将其磷酸基团转移到应答调节蛋白上，使该磷酸化的应答调节蛋白成为阻遏物或诱导蛋白，通过对操纵子的阻遏或激活作用调控下游基因表达。

## F

**翻译 (translation)**：指将 mRNA 链上的核苷酸从一个特定的起始位点开始，按每 3 个核苷酸代表一个氨基酸的原则，依次合成一条多肽链的过程。

**翻译后转运机制(post-translational translocation)**：蛋白质从核糖体上释放后才发生转运。

**翻译-转运同步机制(co-translational translocation)**：某个蛋白质的合成和转运是同时发生的。

**泛素、泛蛋白(ubiquitin)**：含有高度保守的 76 个氨基酸的序列，它以羧基基团连接到目标蛋白质的赖氨酸残基的 $\varepsilon$ 位氨基上，其主要作用是起始蛋白质的降解。

**反式作用因子**:是指能直接或间接地识别或结合在各类顺式作用元件核心序列上参与调控靶基因转录效率的蛋白质。

**非核苷酸型反转录酶抑制剂**:该类化合物可与病毒的反转录酶相结合,通过限制该酶的移动性而影响它的活性,终止病毒 DNA 的合成。

**非组蛋白**:是染色体中除了组蛋白之外的结构蛋白。

**分子伴侣(molecular chaperone)**:它是细胞中一类能够识别并结合到不完全折叠或装配的蛋白质上以帮助这些多肽正确折叠、转运或防止它们聚集的蛋白质,其本身不参与终产物的形成。

**负控诱导**:是原核生物转录调控的一种方式,当阻遏物不与效应物(诱导物)结合时,结构基因不转录。

**负控阻遏**:是原核生物转录调控的一种方式,当阻遏物与效应物(诱导物)结合时,结构基因不转录。

**复制叉(replication fork)**:复制时,双链 DNA 要解开成两股链分别进行 DNA合成,所以,复制起点呈叉子形状,被称为复制叉。

**复制起始点(replication origin)**:是 DNA 链上独特的具有起始 DNA 复制功能的碱基顺序。大肠杆菌的复制起点包括 OriC 和 OriH,OriC 是首选的复制起点,而 OriH 是在 RNaseH 缺失突变株中发现的一系列复制起点。

**复制子(replicon)**:单独复制的一个 DNA 单元被称为一个复制子,它是一个可移动的单位。一个复制子在任何一个细胞周期只复制一次。

## G

**冈崎片段(Okazaki fragment)**:是在 DNA 半不连续复制中产生的长度为 1 000~2 000 个碱基的短的 DNA 片段,能被连接形成一条完整的 DNA 链。

**高速泳动蛋白(high mobility group protein,HMG 蛋白)**:是一类能用低盐溶液抽提、能溶于 2%的三氯乙酸、相对分子质量在 $3.0\times10^4$ 以下的非组蛋白。因其相对分子质量小、在凝胶电泳中迁移速度快而得名。分为 HMG1 和 HMG2 两大类。这类蛋白的特点是能与 DNA 结合,也能与 $H_1$ 作用,但与 DNA 的结合并不牢固,可能与 DNA 的超螺旋结构有关,在 DNA 复制和重组中起重要作用。

**果蝇**:一种双翅目昆虫,其幼虫和成虫依赖于正在腐烂的果实。个体小、生命周期快、繁殖容易,每只雌虫两周就可以产生约 300 个后代。此外,果蝇细胞中的巨大多线染色体使它非常适合于遗传分析和基因定位,是基因与发育领域最重要的模式生物之一。

## H

**核定位序列(nuclear localization sequence,NLS)**:蛋白质中的一种常见的结构域,通常为一短的氨基酸序列,它能与入核载体相互作用,将蛋白质运进细胞核内。

**核苷酸型反转录酶抑制剂**:其结构与脱氧核苷酸类似,多为双脱氧核苷衍生物,可与细胞内自由核苷竞争性地结合反转录酶,终止反转录反应,使病毒

DNA 的合成受阻。

**核酶(ribozyme)**:是一类具有催化活性的 RNA 分子,通过催化靶位点 RNA 链中磷酸二酯键的断裂,特异性地剪切底物 RNA 分子,从而阻断基因的表达。

**核小体(nucleosome)**:是染色质的基本结构单位,由大约 200 bp 的 DNA 和组蛋白八聚体及外围 $H_1$ 蛋白所组成。

**后随链(lagging strand)**:在 DNA 复制过程中,与复制叉运动方向相反的方向不连续延伸的 DNA 链被称为后随链或滞后链。

**花分生组织决定基因**:这类基因促进从花序分生组织产生花分生组织并进一步分化产生花器官原基,但产生何种花器官则由同源域基因所控制。花分生组织决定基因可以在一定程度上激活同源域基因。

## J

**基因(gene)**:产生一条多肽链或功能 RNA 所需的全部核苷酸序列。

**基因表达调控(gene regulation)**:所有生物的遗传信息,都是以基因的形式储存在细胞内的 DNA(或 RNA)分子中,随着个体的发育,DNA 分子能有序地将其所承载的遗传信息,通过密码子-反密码子系统,转变成蛋白质或功能 RNA 分子,执行各种生理生物化学功能。这个从 DNA 到蛋白质或功能 RNA 的过程被称为基因表达,对这个过程的调节就称为基因表达的调控。

**基因表达系列分析技术(serial analysis of gene expression,SAGE)**:是一种以 DNA 序列测定为基础定量分析全基因组表达模式的技术,能够直接读出任何一种细胞类型或组织的基因表达信息。在转录组水平上,任何长度超过 9~10 个碱基的核苷酸片段都可能代表一种特异性的转录产物,因此,用特定限制性核酸内切酶分离转录产物中具有基因特异性的 9~10 个碱基的核苷酸序列并制成标签,将这些序列标签连接、克隆、测序后,根据其占总标签数的比例即可分析其对应编码基因的表达频率。

**基因捕获(gene trapping)**:是检测动植物细胞中基因表达和基因功能的手段。向某个基因组中系统性随机导入带有报道基因的 DNA 片段,就可能在破坏靶基因功能(获得新的表现型)的同时,了解靶基因的表达模式,确定被破坏基因的位置,为进一步克隆靶基因奠定物质基础。

**基因重排**:通过基因的转座或 DNA 的断裂错接等形式使正常基因序列发生改变,使一个基因从远离启动子的地方移到距它较近的位点从而启动转录。

**基因定点突变(site-directed mutagenesis)**:向靶 DNA 片段中引入所需的变化,包括碱基的添加、删除或改变,是分子生物学研究中一种非常有用的手段。

**基因家族**:在基因组进化中,一个基因通过基因重复产生了两个或更多的拷贝,这些基因即构成一个基因家族,是具有显著相似性的一组基因,编码相似的蛋白质产物。

**基因克隆(gene cloning)**:在分子生物学上,人们把将外源 DNA 插入具有复制能力的载体 DNA 中,转入宿主细胞使之得以永久保存和复制的过程称为基因克隆。基因工程或重组 DNA 技术则侧重于验证上述过程所获得遗传物质

新组合在宿主细胞内的表达与功能鉴定。

**基因领域效应**:当两个基因相距太近时,往往不易形成有利于高效转录的空间结构。因此,基因与基因之间的间隔距离被定义为"基因领域"。同一 DNA 链上两个具有相同转录方向的基因间隔小于一定长度时,影响有效转录所必需的染色质结构的形成,从而使这两个基因中的一个或两个均不能转录或转录活性显著降低,既产生了所谓的"基因领域效应"。

**基因敲除(gene knock-out)技术**:针对一个序列已知但功能未知的基因,从 DNA 水平上设计实验,彻底破坏该基因的功能或消除其表达机制,从而推测该基因的生物学功能。

**基因组 DNA 文库(cDNA/genomic DNA library)**:是某一生物体全部或部分基因的集合。将某个生物的基因组 DNA 或 cDNA 片段与适当载体在体外重组后,转化宿主细胞,所得的菌落或噬菌体的集合即为该生物的基因组 DNA 或 cDNA 文库。

**基因芯片(DNA microarray)技术**:把大量已知或未知序列的 DNA 片段点在尼龙膜或玻璃片上,再经过物理吸附作用达到固定化。也可以直接在玻璃或金属表面进行化学合成,得到寡聚核苷酸芯片。将芯片与待研究的 cDNA 或其他样品杂交,经过计算机扫描和数据处理,便可以观察到成千上万个基因在不同组织或同一组织不同发育时期或不同生理条件下的表达模式。

**基因型(genotype)与基因分型(genotyping)**:除性染色体外,人类细胞内的染色体都有两份,因此,一个人所拥有的一对等位位点的类型被称作基因型。事实上,基因型也可以泛指同一基因座位上多个等位位点的类型。确定某个体基因型的实验就称为基因分型。

**基因治疗**:基因治疗是将具有治疗价值的基因,即"治疗基因"装配于带有在人体细胞中表达所必备元件的载体中,导入人体细胞,通过靶基因的表达来治疗遗传疾病。基因治疗是从根本上治疗遗传病的唯一途径。目前科学界关注的主要问题是基因治疗的有效性、安全性和质量可控性。

**基因组(genome)**:生物有机体的单倍体细胞中的所有 DNA,包括核中的染色体 DNA 和线粒体、叶绿体等亚细胞器中的 DNA。

**间隙基因**:参与果蝇体节精细结构形成的基因,该基因最初在整个胚胎中都有很弱的表达,以后随着卵裂的进行而逐渐转变成一些不连续的表达区带。

**简并(degeneracy)**:由一种以上密码子编码同一个氨基酸的现象,对应于同一氨基酸的密码子称为同义密码子(synonymous codon)。

**浆细胞**:即抗体分泌细胞,是 B 淋巴细胞在抗原刺激下分化增殖而形成的一种不再具有分化增殖能力的终末细胞。在分化过程中获得特有的浆细胞抗原,这是浆细胞区别于淋巴细胞的主要标志,可合成并分泌抗体。

**酵母人工染色体(YAC)**:迄今为止容量最大的克隆载体,插入片段平均长度为 200~1 000 kb,最大可达 2 Mb,用于构建基因组文库或物理图谱。

**校对(proofreading)**:在复制、转录或翻译过程中校正错误的机制,包括从正在伸长的核酸或蛋白质链中去除不正确掺入的核苷酸或氨基酸,并以正确的

单元取代。

**聚合酶链反应(polymerase chain reaction,PCR)**:是指通过模拟体内 DNA 复制方式在体外选择性地将 DNA 某个特定区域扩增出来的技术。

## K

**抗体(antibody)**:由特异性免疫途径中的 B 淋巴细胞合成的免疫球蛋白,能够识别抗原上的特定位点。

**抗原 (antigen)**: 指进入人体后能与淋巴细胞产生免疫反应的任何体外物质,包括蛋白或核酸等生物大分子。病原体进入人体后也是抗原。

**抗原呈递细胞**:是免疫应答起始阶段的重要辅佐细胞,有多种类型,可有效摄取、加工和呈递可溶性抗原。其中巨噬细胞分布最广,是处理抗原的主要细胞。

**抗终止因子(anti-termination factor)**:能够在特定位点阻止转录终止的一类蛋白质。它们能与 RNA 聚合酶结合,帮助 RNA 聚合酶越过具有茎环结构的终止子继续转录目标 RNA。

**可译框架、可读框(open reading frame,ORF)**:是指一组连续的含有三联密码子的能够被翻译成为多肽链的 DNA 序列。它由起始密码子开始,到终止密码子结束。

## L

**厘摩(centimorgan)**:重组频率的测量单位。1 厘摩相当于在一个世代中,由于交换而使一个遗传位点标记从第二个遗传位点标记中分离出来的可能性是 1%。在人类基因组中,1 厘摩大约相当于 100 万个碱基对。

**淋巴细胞**:是白细胞的一种,约占白细胞总数的 1/4,直径 6~18 μm,但绝大部分为 6~9 μm 的小淋巴细胞,核圆形,胞浆极少,在免疫反应过程中起重要作用。按其在体内不同的发育成熟途径,可分为 T、B 淋巴细胞两大类。其中依赖于胸腺的称 T 淋巴细胞,具有细胞免疫功能,另一类不直接依赖于胸腺的称 B 淋巴细胞,具有体液免疫功能。

**轮性**:金鱼草、拟南芥等植物的花都由四种类型的花器官组成,花器官排列成向心的圆环形,称作轮性(whorl)。

## M

**孟德尔(Gregor Mendor)**:奥地利修道士,经典遗传学创始人,他的豌豆杂交试验可能是遗传学历史上最具有传奇色彩并且最引人入胜的研究成果,是遗传学的奠基石。

**免疫共沉淀(CO-Immunoprecipitation,CO-IP)**:当细胞在非变性条件下被裂解时,完整细胞内存在的许多蛋白质-蛋白质间的相互作用被保留下来。如果用蛋白质 X 的抗体免疫沉淀 X, 那么在体内与 X 结合的蛋白质 Y 也能被沉淀下来。

**免疫球蛋白**:指动物体内具有抗体活性的蛋白质,主要存在于血浆中,也可

见于其他体液、组织和一些分泌液中。人血浆内的免疫球蛋白大多数存在于丙种球蛋白(γ-球蛋白)中。可分为五类,其中 IgG 是最主要的免疫球蛋白,IgG 分子由 4 条肽链组成。其中分子量为 2.5 万的肽链称为轻链,分子量为 5 万的肽链称为重链。轻链与重链之间通过二硫键相连接。免疫球蛋白是机体受抗原(如病原体)刺激后产生的,其主要作用是与抗原发生免疫反应,生成抗原-抗体复合物,从而阻断病原体对机体的危害,使病原体失去致病作用。免疫球蛋白有时也有致病作用,临床上的过敏症状如花粉引起的支气管痉挛等就是由免疫球蛋白引起的。

**模板链(template strand)**:指 DNA 双链中能作为转录模板通过碱基互补原则指导 mRNA 前体合成的 DNA 链,又称反义链(antisense strand)。

**摩尔根(Thomas Morgan)**:美国著名的遗传学家,研究果蝇的可遗传突变机制,首次证实"基因学说"。

**魔斑核苷酸(magic spot nucleotide)**:受严紧控制的细菌生长过程中一旦缺乏氨基酸供应,细菌会产生一个应急反应,使蛋白质和 RNA 的合成速率迅速降下来。魔斑核苷酸指的就是此过程中由大量 GTP 合成的鸟苷四磷酸(ppGpp)和鸟苷五磷酸(pppGpp),它们的主要作用可能是影响 RNA 聚合酶与启动子结合的专一性,诱发应急反应,帮助细菌渡过难关。

**母源影响基因**:果蝇卵子发育时,卵母细胞自身的细胞核不具转录活性,而由母源抚育细胞、滤泡细胞和脂肪体细胞利用自身的基因和细胞资源提供遗传信息和营养物质,然后输入到卵母细胞中。这些被输入到卵母细胞的基因被统称为母源影响基因。

## N

**内含子的变位剪接**:在高等真核生物中,内含子通常是有序或组成性地从 mRNA 前体中被剪接。然而,在个体发育或细胞分化的某个或某些特定阶段可以有选择性地越过某些外显子或某个剪接点进行 RNA 剪接,产生出组织或发育阶段特异性 mRNA,称为内含子的变位剪接。

**凝胶滞缓实验(DNA mobility shift assay,EMSA)**:是一种检测蛋白质和 DNA 序列相互结合的技术,其基本原理是蛋白质可以与末端标记的核酸探针结合,电泳时这种复合物比没有蛋白结合的探针在凝胶中泳动速度慢,表现为相对滞后。该方法可用于检测 DNA 结合蛋白、RNA 结合蛋白,并可通过加入特异性的抗体来检测特定的蛋白质。

## Q

**启动子(promoter)**:与基因表达启动相关的顺式作用元件,是结构基因的重要成分。它是一段位于转录起始位点 5′端上游区大约 100~200bp 以内的具有独立功能的 DNA 序列,能活化 RNA 聚合酶,使之与模板DNA 准确地相结合并具有转录起始的特异性。

**前导链(leading strand)**:在 DNA 复制过程中,与复制叉运动方向相同,以

5′-3′方向连续合成的链被称为前导链。

**切除修复(excision repair)**:DNA 损伤的一种修复机制,直接切除受损伤的一条 DNA 片段,以其互补链为模板新合成 DNA 来取代切除的受损片段。

**禽流感病毒 H5N1 型**:A 型流感病毒根据其表面血凝素(H)和神经氨酸酶(N)结构的不同可分为许多亚型:血凝素(H)有 15 个亚型(H1-H15),神经氨酸酶(N)有 9 个亚型(N1-N9),它们之间的不同组合,使 A 型流感病毒出现许多亚型(如 H1N1、H2N2、H5N12 等),理论上可产生 135 种不同的病毒亚型,各亚型之间无交互免疫力。禽流感病毒 H5N1 即具有第 5 亚型血凝素(H)和第 1 亚型神经氨酸酶(N)的流感病毒。

**禽流行性感冒(Avian influenza,AI)**:简称禽流感,又名真鸡瘟、欧洲鸡瘟、鸡疫等,是由 A 型流感病毒(Avian Influenza Virus,AIV)引起的呼吸系统病变直到全身败血症的一种高度急性传染病。鸡、火鸡、鸭等家禽及野鸟均可被感染。禽流感病毒还是人流感病毒株形成的最大的基因库来源。禽流感也可直接感染人,对人类的公共卫生也造成了相当大的危害。

## R

**人类免疫缺陷病毒(HIV)**:俗称艾滋病毒(AIDS),是一种能生存于人的血液中并攻击人体免疫系统的病毒,主要攻击人体免疫系统中重要的 T4 淋巴细胞,大量吞噬、破坏 T4 淋巴细胞,从而使得整个人体免疫系统遭到破坏,最终因丧失对各种疾病的抵抗能力而死亡。

**人禽流感**:又称人禽流行性感冒,是由禽甲型流感病毒某些亚型中的毒株引起的急性呼吸道传染病。

**弱化子(attenuator)**:是指原核生物操纵子中能显著减弱甚至终止转录作用的一段核苷酸序列,该区域能形成不同的二级结构,利用原核生物转录与翻译的偶联机制对转录进行调节。

## S

**上游启动子元件(upstream promoter element,UPE)**:将 TATA 区上游的保守序列称为上游启动子元件或称上游激活序列(upstream activating sequence,UAS)。

**神经氨酸酶(neuraminidase,NA)**:一种糖苷外切酶,是禽流感病毒表面上的两种主要糖蛋白之一,可从 α-糖苷键上除去唾液酸残基,对病毒的释放及病毒在感染细胞周围的扩散能力有很大影响。此外,神经氨酸酶也是禽流感病毒中的重要抗原,它的高突变频率也是禽流感病毒较难防治的原因之一。

**噬菌体(bacteriophage)**:是感染细菌的病毒。烈性噬菌体的感染最终导致宿主细胞裂解,而温和型噬菌体则能将 DNA 整合到宿主细胞染色体上,从而对宿主造成较为长期的影响。在特定条件下两者可以转换。

**噬菌体展示技术(phage display)**:将编码“诱饵”的 DNA 片段插入噬菌体基因组,并使之与噬菌体外壳蛋白编码基因或其他结构基因相融合,用该重组

噬菌体侵染宿主细菌,复制形成大量带有杂合外壳蛋白(或其他结构蛋白)的噬菌体颗粒,从而捕获 cDNA 表达文库中与"诱饵"相互作用的蛋白质。

**顺反子(cistron)**:功能基因,意为通过顺式(基因序列)及反式(所编码的蛋白质)试验所确定的一个遗传学单位。

**顺式作用元件**:存在于基因旁侧序列中能影响基因表达的序列,包括启动子、增强子、调控序列和可诱导元件等,本身不编码任何蛋白质,仅仅提供一个作用位点,与反式作用因子相互作用参与基因表达调控。

## T

**肽酰-tRNA(peptidy-tRNA)**:指在蛋白质生物合成过程中,在肽键合成之后,连接在新生肽链上的 tRNA 分子。

**肽基转移酶(peptidyl transferase)**:蛋白质合成过程中的一种酶,它催化正在延伸的多肽链与下一个氨基酸之间形成肽键。

**体液免疫**:在体液免疫中,B 淋巴细胞首先识别外源致病因子或其他任何形式的抗原,通过浆细胞(即成熟 B 淋巴细胞)分泌出能溶于血液蛋白质中的免疫球蛋白 Ig。这些被称作抗体的 Ig 分子通过特定的结合位点与抗原形成抗原-抗体复合物,最终被巨噬细胞所吞噬。

**同工 tRNA(cognate tRNA)**:指几个代表相同氨基酸、能够被一个特殊的氨酰-tRNA 合成酶识别的 tRNA。

**同源域**:与生物有机体的生长、发育和分化密切相关,广泛存在于真核生物基因组内编码 60 个保守氨基酸序列的 DNA 片段。

**同源域基因**:果蝇发育过程中决定躯体体节命运的基因,躯体部分最终发育成为无翅的前胸还是有翅的中胸,有平衡器的后胸还是腹部体节都由该组基因控制。

## W

**卫星 DNA(satellite DNA)**:又称随体 DNA。因为真核细胞 DNA 的一部分是不被转录的异染色质成分,其碱基组成与主体 DNA 不同,因而可用密度梯度沉降技术如氯化铯梯度离心将它与主体 DNA 分离。卫星 DNA 通常是高度串联重复的 DNA。

**无义突变(nonsense mutation)**:在 DNA 序列中任何导致编码氨基酸的三联密码子转变为终止密码子(UAG、UGA、UAA)的突变,它使蛋白质合成提前终止,合成无功能的或无意义的多肽。

**物理图谱(physical map)**:利用限制性内切酶将染色体切成片段,再根据重叠序列确定片段间连接顺序以及遗传标志之间的物理距离(碱基对,bp)或千碱基对(kb)或兆碱基对(Mb)的图谱。

## X

**系统生物学**:是研究一个生物系统中所有组成成分(基因、mRNA、蛋白质、

糖、次生代谢产物等)的变化规律以及在特定遗传或环境条件下相互关系的学科。系统生物学研究通过整合各组分的信息,以图画或数学方式建立能描述系统结构和行为的模型。

**细胞免疫**:在细胞水平免疫反应中,抗原受体细胞产生类似 Ig 的分子,紧密结合在 T 淋巴细胞膜表面的受体分子上。抗体分子与受体的专一性结合最终导致整个感染细胞被降解。

**细胞因子**:是细胞分泌的具有生物活性的小分子蛋白质的统称。在很多情况下,多种免疫细胞间的相互作用是通过细胞因子介导完成的。细胞因子可有多种名称,如单核巨噬细胞产生单核因子,淋巴细胞产生淋巴因子,可刺激骨髓干细胞或祖细胞分化成熟的细胞因子被称为集落刺激因子。

**细胞转化基因(C-onc)**:即被激活的原癌基因,能够转化 NTH/3T3 成纤维细胞或其他靶细胞成为肿瘤细胞的基因,也称癌基因(oncogene)。

**细菌人工染色体(BAC)**:科学家用细菌的 F 质粒及其调控基因构建的细菌染色体克隆载体,常用来克隆 150 kb 左右大小的 DNA 片段。

**细菌转化(transformation)**:是指一种细菌菌株由于捕获了来自另一种供体菌株的 DNA 而导致性状特征发生遗传改变的过程。

**小分子干扰核糖核酸 (siRNA,short interfering RNAs)**: 是长度为 21~25 个核苷酸的双链小分子 RNA, 它是引发转录后基因沉默中序列特异性的 RNA 降解的重要中间媒介。siRNA 具有 5′端磷酸基和 3′端羟基,两条链的 3′端各有两个碱基突出于末端。在转录后沉默和 RNAi 的过程中,siRNA 作为识别目标基因的引导物。

**信号识别蛋白(signal recognition protein,SRP)**:由 6 种紧密结合的信号识别蛋白质与一个长约 300 核苷酸的 7S RNA 分子形成的核糖核蛋白颗粒,能识别核糖体上新合成多肽链的前导序列并与其结合,将核糖体、新合成肽链及信号识别颗粒导向内质网,使肽链的翻译及转运同时进行。

**信号肽(signal peptide)**:在起始密码子后,有一段编码疏水性氨基酸序列的RNA 区域,被称为信号肽序列,它负责把蛋白质引导到细胞内不同膜结构的亚细胞器内。

**序列标签位点(sequence tagged site)**:序列标记位点(STSs)指基因组中物理位置已被确定的小段单拷贝序列,是功能基因在染色体上位置的标记。

**血凝素**:即红血球凝聚素,是流感病毒表面的一类蛋白质,具有很高的突变率,使病毒能逃避宿主的免疫系统并具有抗药性。病毒依靠它与宿主细胞表面相结合。

## Y

**阳离子多聚体**:是一种表面具有带正电荷的氨基基团的多聚体,氨基基团可与 DNA 的磷酸基团结合发生电性中和,将 DNA 吸附而使其不易被核酸酶降解,并可防止沉淀,从而提高转染效率。

**野生型和突变型(wild-type and mutant)**:生物学上把所研究的基因位点未发生改变的生物个体称为野生型,把所研究的基因位点发生遗传突变的生物

个体称为突变型。因此,野生型和突变型是相对的,甚至可能根据所研究目的基因的不同而发生转换。

**移码突变(frameshift mutation)**:指一种突变,其结果可导致核苷酸序列与相对应蛋白质的氨基酸序列之间的正常关系发生改变。移码突变是由删去或插入一个核苷酸的"点突变"构成的,突变位点之前的密码子不发生改变,但突变位点以后的所有密码子都发生变化,编码的氨基酸出现错误。

**移植抗原**:在不同种属或同种不同系的动物个体间进行正常组织移植时会出现排斥,是供者与受者组织不相容的反映。排斥反应本质上是一种免疫反应,受组织表面的同种异型抗原诱导,这种代表个体特异性的同种抗原称为组织相容性抗原或移植抗原。

**遗传密码(codon)**:mRNA 上每 3 个核苷酸翻译成多肽链上的一个氨基酸,这 3 个核苷酸就称为一个密码子(三联子密码)。

**遗传图谱(genetic map)**:某一物种的染色体图谱(或称连锁图谱),显示全体已知基因和/或遗传标记的相对位置,而不是它们在每条染色体上特殊的物理位置。

**乙肝病毒(HBV)**:是嗜肝 DNA 病毒科中哺乳动物病毒属的一员。完整的乙型肝炎病毒颗粒直径为 42nm,又名 Dane 颗粒,具有双层核壳结构,外壳相当于包膜,含有乙型肝炎病毒表面抗原和多聚人血清白蛋白受体等。内部为直径 28nm 的核心颗粒,核心颗粒表面含有乙型肝炎病毒核心抗原(HBcAg)和乙型肝炎 e 抗原(HBeAg),颗粒内部有乙型肝炎病毒的 DNA 和DNA 聚合酶。

**引发酶(primase)**:是依赖于 DNA 的 RNA 聚合酶,其功能是在 DNA 复制过程中合成 RNA 引物。

**引发体(primosome)**:DNA 复制过程中引发合成每个冈崎片段时所需的多蛋白复合物,包括预引发蛋白、具有 ATP 酶活性的蛋白质以及引物酶。引发体与 DNA 结合后由引物酶合成 RNA 引物并合成与 RNA 引物相连接的冈崎片段。引发体沿不连续合成的 DNA 链移动,其移动的方向与 RNA 及 DNA 的合成方向相反。引发体移动需要来自 ATP 水解的能量。

**引物(primer)**:是指一段较短的单链 RNA 或 DNA,它能与 DNA 的一条链配对提供游离的 3′-OH 末端以作为 DNA 聚合酶合成脱氧核苷酸链的起始点。

**原病毒**:逆转录病毒侵入宿主细胞后,首先利用自身携带的逆转录酶合成出与本身 RNA 基因组互补的 DNA,再利用宿主细胞的 DNA 聚合酶指导合成出另一 DNA 链,以双链 DNA 形式整合到宿主细胞基因组中。被整合的反转录病毒 DNA 分子就称为原病毒。

**原位杂交(*in situ* hybridization,ISH)**:是用标记的核酸探针,经放射自显影或非放射检测体系,在组织、细胞及染色体水平上对核酸进行定位和相对定量研究的一种手段。通常分为 RNA 原位杂交和染色体原位杂交两大类。

## Z

**载体(vector)**:能将外源 DNA 或基因片段携带入宿主细胞内的一个具有自

主复制能力的 DNA 分子。

**增强子(enhancer)**:能提高转录起始效率的序列被称为增强子或强化子。增强子可位于转录起始点的 5′或 3′末端,而且一般与所调控的靶基因的距离无关。

**脂质体**:脂质体是具有双层膜的封闭式粒子,它们能促进极性大分子穿透细胞膜。根据脂质体包裹 DNA 的方式不同可将脂质体分为阳离子脂质体、阴离子脂质体、pH 敏感脂质体及融合脂质体等。

**指导 RNA(guide RNA)**:原生动物及植物线粒体中进行 RNA 编辑所需的 RNA 序列,是与已正确编辑的 RNA 序列互补的一小段 RNA,被用来作为向未经编辑的 RNA 中插入碱基的模板。

**中心法则(central dogma)**:由克连克首次提出的遗传信息传递规律,该法则阐明了 DNA 复制、RNA 转录以及翻译产生蛋白质在生命过程中的核心地位。

**种系类型**:在生殖细胞或不表达抗体基因的体细胞中,免疫球蛋白的 V 基因和 C 基因在染色体上呈分隔独立排列的结构,这种排列类型被称为种系类型。在有效表达抗体基因的免疫细胞中发生重排。

**转录单元(transcription unit)**:是一段可被 RNA 聚合酶转录成一条连续 mRNA 链的 DNA,包括转录起始和终止信号。一个简单的转录单位只携带合成一种蛋白的信息,复合转录单位可携带不止一种蛋白质分子的信息。

**转录起始位点(transcription initiation site)**:是指与新生 RNA 链第一个核苷酸相对应 DNA 链上的碱基位点,通常为嘌呤。常把起点前,即 5′末端的序列称为上游(upstream),而把其后面即 3′末端的序列称为下游(downstream)。

**转座、移位(transposition)**:遗传信息从一个基因座转移至另一个基因座的现象称为基因转座,是由可移位因子(transposable element)介导的遗传物质重排。

**转座子(transposon,Tn)**:是存在于染色体 DNA 上可自主复制和位移的基本单位。参与转座子易位及 DNA 链整合的酶称为转座酶。

**自主复制序列 (autonomously replicating sequences,ARS)**:是酵母 DNA 复制的起点,长约 150 bp 左右,包括数个复制起始必需的保守区。不同 ARS 序列的共同特征是有一个被称为 A 区的 11bp 的保守序列。

**阻遏蛋白(repressor)**:是指转录调控系统中调节基因表达产物丰度的蛋白质,其作用部位往往是操纵子的操纵区,起着阻止结构基因转录的作用。

**组蛋白(histones)**:是保守的 DNA 结合蛋白,是染色体的结构蛋白,分为 H1、H2A、H2B、H3 及 H4 五种,与 DNA 共同组成真核生物染色质的基本单位核小体。

**组织相容性复合体(MHC)**:能引起强而迅速的排斥反应的抗原称为主要组织相容性抗原,其编码基因是一组紧密连锁的基因群,称为主要组织相容性复合体。

# 主要参考书目


1. 吴乃虎编著. 基因工程原理. 第二版. 北京:科学出版社,2001
2. 瞿礼嘉,顾红雅,胡苹,陈章良编著. 现代生物技术. 北京:高等教育出版社,2004
3. 张红卫,王子任,张士璀编著. 发育生物学. 第二版. 北京:高等教育出版社,2006
4. Brown T A. Genomes. 3rd ed. New York:John Wiley & Sons Inc. Ltd,2006
5. Buchanan B B,Gruissem W,Jones R L. Biochemistry & Molecular of Plants. Maryland:American Society of Plant Physiologists,2000
6. Lewin B. Genes Ⅷ. Pearson Education,Inc.,2004
7. Lodish H,Baltimore D,Berk A,Zipursky S L,Matsudaira P,Darnell J. Molecular Cell Biology. 5th ed. New York:Scientific American Books,2003
8. Nelson D L,Cox M M. Lehninger Principles of Biochemistry. 4th ed. New York:Worth Publishers,2004
9. Watson J D,Baker T A,Bell S P,Gann A,Levine M,Losick R. Molecular Biology of the Gene. 5th ed. Cold Spring Harbor Laboratory Press,Pearson Education,Inc.,2004

# 第三版后记

在本书的编写过程中，得到北京大学蛋白质工程及植物基因工程国家重点实验室、生命科学学院生物化学及分子生物学系陈章良、顾孝诚、罗静初、朱圣庚、张庭芳、潘乃穆等许多老师的帮助和关心，在此表示衷心的感谢。书中的不少插图引自《Genes Ⅷ》、《Molecular Biology of the Cell》、《Molecular Cell Biology》和《Plant Virology》，在此向原书作者 B.Lewin，H.Lodish，D.Baltimore，A.Berk，S.L.Zipursky，P.Matsudaira，J.Darnell，B.Alberts，D.Bray，J.Lewis，M.Raff，K.Roberts，J.Watson，R.E.F.Matthews 表示感谢。王莉、苏彦辉、任晓菲、敬群、李小华、梁西卉、唐家福、周彩芬、郑美兰、魏春红等同志及朱玉贤实验室的不少高年级博士研究生参与资料搜集、打印和图表处理等工作，在此一并致谢。

愿我们的拙作能为广大青年学子的成长贡献绵薄之力，敬请全国同行批评指正。

编　者

2007 年 2 月，北京

# 索　引


5-甲基胞嘧啶　293
7-甲基鸟嘌呤　293
Ⅰ类内含子　98
Ⅱ类内含子　98
α-鹅膏蕈碱　71
α 亚基　70
β-半乳糖苷酶　231,237
β-半乳糖苷透过酶　237
β-半乳糖苷乙酰基转移酶　237
β-半乳糖酶　144
β-珠蛋白基因　79,90
θ 型　41
κ 基因　391
λ 基因　391
λ 噬菌体　92
MHC　397
ρ 因子　91,451
σ 因子　69,70,234,451


## A


ABC 模型　409,448
A-DNA　33
Ac-Ds 系统　59
ADP 核糖基化　24
Altman　10
AP1 家族　306

AP 位点 53,448
AraC 蛋白 259
ara 操纵子 258
Avery 6,10
A 位点 131
阿拉伯糖 257
阿拉伯糖异构酶 257
癌 339,451
癌基因 338
艾滋病(AIDS)病毒 351
安慰性诱导物 239,451
氨酰-tRNA 合成酶 107,120,451

## B

B-DNA 33
BAC 418
Baltimore 9,156
Berg 9,156
Bishop 10
Blobel 10
Boyer 156
Brener 10
Brenner 156
Brinster 156
Burkitt 淋巴瘤 349
B 细胞 386
靶标鸟枪法 421
摆动假说 115,451
半保留复制 37,38
半不连续复制 46
半胱氨酸 119
半乳糖-磷酸尿嘧啶核苷转移酶 255
半乳糖操纵子 255
半乳糖激酶 255
胞嘧啶 18,32
比较基因组学 443,451
边界序列 287
编码链 66,451
表达序列标签 451
表达序列标签图 433
表面抗原主蛋白 367
病毒癌基因 338,451
病毒载体 378,451
泊松公式 419
不可逆调控 281
布朗运动 76

## C

CAAT 盒 297
CAAT 区 77
cAMP 242
CDK2 315
cDNA 172,448
Cech 10
*chal* 431
Cohen 158
ColE1 50
contigs 431
Copia 60
CpG 岛 292
Crick 6,9,10,156
crp⁻ 256
cya⁻ 256
*C-ras* 基因 346
C 值 25,448
C 值反常现象 25,448
蚕丝心蛋白基因 332
操纵区 240
操纵子 9,84,237,451
插入序列 57,452
差向异构酶 257
长末端重复序列 347,452
超螺旋 36,452
超螺旋数 37
成对规则基因 405,452
重叠基因 31,276,452
重叠延伸 198
重链基因的重排 393
重组率 187
重组种系理论 392,452
传感蛋白 262
雌激素 324
次级氢键 118
从头合成型甲基转移酶 292
催乳素 334
错配修复 52,452

错义突变 119,452

## D

D-环型 41
Dane 颗粒 364,448
DBRP 316
DNA 448
DnaQ 蛋白 263
DNA 的半保留复制 38,449
DNA 的半不连续复制 46,449
DNA 甲基化 292,449
DNA 结合蛋白 25
DNA 聚合酶 47,449
DNA 酶 I 超敏感位点 289,449
DNA 修复系统 260
DNA 拓扑异构酶 45,449
Doherty 10
Dulbecco 9
DuPont 公司 157
D 臂 115
达尔文 2
大肠杆菌 433,438
大肠杆菌基因组 230,438
大鼠降钙素基因 331
大亚基 123
代谢物激活蛋白 242
代谢物阻遏效应 242,452
单倍体 22
单倍型 452
单核苷酸的多态性 430,452
单顺反子 mRNA 84,452
蛋白酶抑制剂 363
蛋白质 126
蛋白激酶 308,452
蛋白质磷酸化 226,307,452
蛋白质芯片 209
蛋白质乙酰化 316,453
蛋白质印迹法 189
蛋白质转运 142
蛋白质组与蛋白质组学 187,453
倒置重复区 57
"灯刷型"染色体 289
等离子表面共振技术 209
等位基因 423,453
等位排斥与同型排斥 396,453
等位位点 178,453
第二遗传密码 115
颠换 430
电泳的迁移率 161
淀粉酶基因 287
定点突变 197
豆血红蛋白 268
断裂基因 286
多聚 A 位点(poly A 位点) 453
多聚阴离子 161
多顺反子 30
多顺反子 mRNA 30,84,453
多余臂 116

## E

*ex vivo* 途径 377
E 位点 131
二倍体 22
二硫键 136
二氢叶酸还原酶基因 286
二乙基己烯雌酚 324
二元闭合复合物 78
二元开链复合物 78
二组分系统 262,453

## F

Far Western 208
FeMoCo 264
Fiers 31
Fire 10
FRET 212
翻译 66,107,453
翻译-转运同步机制 143,453
翻译后转运机制 143,453
翻译偶联 276
反密码子 115
反密码子臂 115
反式作用因子 296,454
反义链 67
泛素、泛蛋白 453
非编码区 84

非复制型 62
非核苷酸型反转录酶抑制剂 363,454
非洲爪蟾 158
非自主性因子 59
非组蛋白 22,454
肺炎链球菌 6
肺炎支原体 433
分离规律 5
分裂间期 21,27
分枝酸 246
分子伴侣 140,454
分子信标技术 180
封闭复合物 73
辅阻遏蛋白 246
负超螺旋 36
负控诱导 232,454
负控阻遏 232,454
负转录调控 232
附加受体 399
复合型转座子 57
复杂多基因家族 284
复制叉 38,454
复制起始点 38,454
复制期 51
复制型 62
复制预备期 51
复制子 38,454

## G

*GAL4* 306
GC 盒 297
Gilbert 9,156
Gilman 10
*glk1* 431
*glnA* 基因 267
GlnB 蛋白 267
*glnD* 基因 267
GMO 156
Griffith 6
GST 融合蛋白沉降技术 209
GU-AG 法则 95,449
高速泳动蛋白 24,454
干扰素基因 286
肝脏 287
冈崎片段 46,454
根瘤素 268
供体位点 287
共聚物 110
共受体 399
共线 434
狗基因组 157
固醇类激素 322
固氮菌 263
固氮酶 264
寡聚(dT)-纤维素柱 173
冠状病毒 373
光解酶 55
光纤 216
滚环型 41
果蝇 400,433,454

## H

Hartwell 10
*Ha-ras* 基因 346
Higa 158
HLA 398,449
HMG 蛋白 24
Hogness 区 77
Holley 156
Holly 9
HOM 复合体 406
Hooke 3
Hunt 10
海胆 284
合子 22
核定位序列 150,454
核苷酸 18
核苷酸切除修复 53
核苷酸型反转录酶抑制剂 363,454
核仁 RNA 102
核酶 102,455
核糖 18
核糖体 107,121,230
核糖体结合技术 110
核酮糖激酶 257
核小体 27,455

核心酶 69
核心启动子 297
核运转因子 150
黑猩猩基因组 157
恒定区 290
后随链 44,455
化学修饰 329
环腺苷酸 236
环腺苷酸受体蛋白 236
获得性免疫缺陷综合征 351
获得性遗传 105
花分生组织决定基因 407,455

## I

$Ig^{0}$,$Ig^{+}$,$Ig^{-}$ 449
*in vivo* 途径 378
IPTG 242
IRE 334
IRE 结合蛋白 334
Iwoff 9

## J

Jacob 9,156
Jerne 10
Jun 306
肌红蛋白 9
肌红蛋白重链基因 287
肌球蛋白碱性轻链基因 331
基因 5,66,87,297,455
基因表达调控 13,231,455
基因表达系列分析技术 193,455
基因工程 12,155,449
基因定点突变 197,455
基因家族 282,455
基因扩增 290
基因领域效应 349,456
基因捕获 204,455
基因敲除技术 200,456
基因克隆 455
基因芯片技术 214,456
基因治疗 377,456
基因型 179,456
基因重排 290,347
基因组 416,456
基因组 DNA 文库 170,456
基因组学 423
激活蛋白 234
甲基化 24
甲基化鸟嘌呤 87
甲硫氨酸 119
甲酰基转移酶 128
甲酰甲硫氨酸 119
间隙基因 405,456
简并 111,456
简单多基因家族 282
简单性状 4
碱基 18,32
碱基互补配对 33
碱性-亮氨酸拉链 303
碱性-螺旋-环-螺旋 304
浆细胞 386,456
降解物敏感型操纵子 243
降解物抑制作用 236
交叉点 289
交配型转换 291
胶原蛋白基因 286,289
焦磷酸转移酶 236,278
酵母单杂交 206
酵母基因组 157
酵母人工染色体 418,456
酵母双杂交 207
校对 49,456
校正 tRNA 120
接合体 119
结合核糖体 144
解离酶 62
进化论 2
茎-环结构 92,249
精氨酸 23
警报素 278
巨噬细胞 399
聚丙烯酰胺 161
聚合酶链反应 165,457
均聚物 110

## K

Kelley 156
Kendrew 9
Khorana 156
*Ki-ras* 346
Klenow 大片段 418
Kohler 10
Kornberg 9,10,156
Kossel 9
卡那霉素 141
开放复合物 73
抗体 177,384,457
抗原 177,384,457
抗解链蛋白 244
抗原呈递细胞 387,457
抗终止因子 271,457
可变区 290
可逆性调控 281
可译框架、可读框 108,457
可诱导基因 235
可阻遏基因 235
克隆 182
克隆羊 157
控制因子 59
枯草芽胞杆菌 433
跨域作用因子 296

## L

L-核酮糖-5-磷酸-4-差向异构酶 257
*lac* 操纵子 240
Leeuwenhoek 3
Lewis 10
LexA 阻遏蛋白 260
LINEs 429
LTR 347,357
L 形折叠 117
赖氨酸 23
劳斯(Rous)肉瘤病毒 341
酪蛋白 mRNA 334
类菌体 269
厘摩 187,425,457
连接区 290
连接数 37
连锁图 425
连锁遗传 5
链霉素 141
两步置换法 221
亮氨酸 121
烈性噬菌体 225
邻氨基苯甲酸合酶 246
邻氨基苯甲酸磷酸核糖转移酶 246
邻位互调假说 325
淋巴细胞 384,457
磷酸 18,32
磷酸二酯键 34,69
磷酸化 24
流感嗜血杆菌 433
氯高铁血红素 334
氯化钙 158
氯霉素 140
卵壳蛋白 290
卵裂 401
卵母细胞 290
轮性 457
螺线管 29
螺旋-转折-螺旋 300

## M

Mandel 158
MADS-box 409,449
Maxam 156
Maxam-Gilbert 417
McClintock 9,57,59
Mello 10
Meselson 10
MHC 397
Miescher 156
Milstein 10
Monod 9,156
mRNA 67,83
Mullis 10
麦芽糖 144
脉冲电场凝胶电泳 162
锚定酶 194
孟德尔 4,457

免疫 10
免疫共沉淀 209,457
免疫球蛋白 389,457
免疫球蛋白结构基因 290
免疫球蛋白重链基因 391
免疫系统 384
模板链 66,458
摩尔根 5,458
魔斑核苷酸 277,458
母源影响基因 401,458

## N

Nef 蛋白 361
*nifAL* 267
*nifD* 264
*nifH* 264
*nifK* 264
Nirenberg 9,156
Notch 信号途径 388
NTPs 67
*ntrA* 267
*ntrB* 267
*ntrC* 267
Nurse 10
Nusslein-Volhard 10
$N^6$-甲基腺嘌呤 292
N 区 394,449
内含子 93,286
内含子的变位剪接 97,458
拟南芥 157,433
酿酒酵母 218,433
鸟苷四磷酸 236
鸟苷五磷酸 236
鸟嘌呤 18,32
尿苷二磷酸半乳糖 257
尿苷酸转移酶 267
尿殖道支原体 433
凝胶滞缓实验 223,458
柠檬酸合成酶基因 446
诺贝尔 9

## O

*O*-硝基半乳糖苷 238
OriC 50
OriH 50

## P

*p53* 基因 315,350
Palmiter 156
Perutz 9
pRb 315
Pribnow 区 77
Prusiner 10
*pur* 操纵子 245
p21 蛋白 315
P 位点 131
P 转座子 60,450
盘绕数 37
配子 22
嘌呤霉素 136,141
葡糖-1-磷酸 255
葡萄糖 235

## Q

启动子 67,75,234,298,458
起始 tRNA 119
前导链 44,458
前导区 250
前导肽 147,249
亲核基团 98
禽流感病毒 370,459
禽流行性感冒 370
切除修复 53,459
琼脂糖 161
巯甲基半乳糖苷 238
全基因组鸟枪法测序 419
全酶 69

## R

*R.Leguminosarum* 264
*R.phaseoli* 264
RACE 183,450
RAPD 185,427,450
*Rb* 基因 350
*recA* 基因 260
Rev 蛋白 359

RFLP 185,426,450
*Rhizobium trifolii* 264
RNase H 175
RNA 的编辑 99,450
RNA 的剪接 93,450
RNA 的再编码 101,450
RNA 干涉(RNAi) 212,450
RNA 聚合酶 69,450
Roberts 10
Rodbell 10
Roderick 423
Rop 蛋白 50
rRNA 67
rRNA、mRNA 107
rRNA 操纵子 262
*rRNA* 基因 290
Rubin 156
染色体 5,19
染色体外遗传因子 30
热激因子 326
热休克蛋白 326
人 α-珠蛋白基因簇 284
人 γ-珠蛋白基因 294
人工染色体技术 418
人类 433
人类基因组 157
人类基因组计划 416
人类免疫缺陷病毒 351,459
人禽流感 369,459
日常型甲基转移酶 292
溶菌周期 225
溶源周期 225
融合基因 246
肉瘤病毒 339
乳糖操纵子 237
朊病毒 10
弱化子 235,459
弱化作用 249

## S

SAGE 193
Sanger 9,31,156,417
SARS-CoV 373,450
Schleiden 3
Schwann 3
SD 序列 85,450
Sharp 10
SINEs 430
SI 操纵子 262
Smith 10
Smith,Wilcox 156
SNP 178,450
Spm-dSpm 系统 60
Spradling 156
Src 同源结构域 313
Stahl 10
STS 技术 431
Sumner 9
三级氢键 118
三元复合物 73
三元转录复合物 91
扫描模型 132
色氨酸合酶 246
上游激活序列 80
上游启动子元件 80,297,459
神经氨酸酶 370,459
生物固氮 263
实时定量 PCR 167
使用频率 276
视网膜纤维瘤 350
噬菌体 6,22,225,459
噬菌体展示技术 225,459
受体臂 115
受体流动假说 324
受体位点 287
枢纽蛋白 341
双螺旋 6,32,118
双向电泳技术 188
双重功能 259
水稻基因组 157
顺反子 460
顺式作用元件 296,460
顺乌头酸酶基因 446
瞬时调控 281
四环素 141

## T

T-DNA 205,450
T-淋巴细胞 388
T-细胞受体基因 290
TATA 区 77,234
Tat 蛋白 358
Temin 9,156
TnA 家族 58
tRNA 67,107,115
*trp* 操纵子 246
*tsx* 基因 245
TTGACA 区 77
$T_{\psi}C$ 臂 115
T 抗原基因 286
T 细胞受体 388
肽键 4
肽基转移酶 132,460
肽酰-tRNA 126,460
糖苷水解酶 53
糖皮质素受体 303
体节极化基因 405
体液免疫 385,460
铁-钼辅助因子 264
铁蛋白 264,334
铁钼蛋白 264
铁应答元件 334
停靠蛋白 145
同工 tRNA 119,460
同型排斥 396
同义密码子 111
同源域蛋白 304
同源域基因 406,409,460
统一律 5
透过酶 240
突变型 5
图位克隆法 185
兔网织红细胞 334
脱氧核糖 32
脱氧核糖核酸 32
拓扑异构酶 47
唾液腺 287

## U

Uchoa 9,156
UDP-葡萄糖 257
UMP 267

## V

Varmus 10
Vif 蛋白 361
Vpr 蛋白 361
Vpu 蛋白 361

## W

Watson 6,9,157,416
Weber 31
Weiner 31
Weissman 31
Wieschaus 10
Wilkins 9
外显子 93,286
卫星 DNA 428,460
微珠芯片技术 216
温和噬菌体 225
无义突变 119,460
五碳糖 257
物理图 431,460

## X

*Xist* 基因 296
X 染色体失活 295
细胞理论 3
细胞免疫 385,461
细胞色素 b 101
细胞学说 8
细胞因子 388,461
细胞转化基因 338,461
细菌人工染色体 420,461
细菌转化 158,163,461
系统生物学 16,460
腺病毒伴随病毒 378
线虫 433
腺苷酸环化酶 236
腺苷酸转移酶 87

腺苷脱氨酶基因　377
腺嘌呤　9,18,32
纤突蛋白 S　373
硝化、脱硝化　263
小分子干扰核糖核酸　213,461
小鼠 α-珠蛋白基因　294
小鼠淀粉酶基因　287
小鼠乳腺癌　349
小卫星 DNA　428
小亚基　123
效应物　232
缬氨酸　121
锌指　301
新霉素　141
信号识别蛋白　145,461
信号肽　144,461
胸腺嘧啶　9,18,32
溴乙啶　162
序列标签位点　431,461
序列图　434
选择性剪接　195,287
血红蛋白　9
血红蛋白基因　284
血凝素　370,461
血凝素酯酶　373

## Y

YAC　418
Yanofsky　10,156
胭脂碱合成酶　90
延伸 tRNA　119
严重急性呼吸系统综合征　373
阳离子多聚体　382,461
野生型和突变型　461
胰岛素　9,156
胰高血糖素　325
移码突变　109,462
移位　56,134
移植抗原　398,462
遗传　6
遗传距离　425
遗传密码　66,108,462
遗传图　425
乙肝病毒　364,462
乙基化　24
异丙基巯基半乳糖苷　238
异构酶　246,255
异构乳糖　240
异柠檬酸脱氢酶基因　446
异硫氰酸胍-苯酚抽提法　172
引发酶　44,45,462
引发体　45,462
引物　165,462
吲哚甘油磷酸合酶　246
应答调节蛋白　262
荧光原位杂交　197
营养细胞　290
有丝分裂期　51
有丝分裂准备期　51
有意义链　66
右剪接位点　287
右手螺旋　33
诱导酶　235
原癌基因　343,346
原病毒　341,462
原核生物　230,280
原位杂交　197,462

## Z

Z-DNA　33
Zinkernagel　10
杂交　4
载体　158,462
载脂蛋白　99
增强子　78,298,463
真核生物　280
正超螺旋　36
正控诱导　234
正控阻遏　234
正向重复区　57
正转录调控　232
直接修复　55
植物凝血素　264
指导 RNA　101,463
脂质体　381,463
中介物假说　324

中心法则 10,99,105,463
终止密码子 111
终止子 76,249
种系类型 391,463
周期蛋白 315
主要组织不相容抗原 10
专利化 156
转化频率 163
转换 430
转基因植物 156
转录 66,118
转录单元 76,463
转录起始位点 76,463
转录起始前复合物 67
转录图 432
转录因子 14
转肽酶中心 126
转铁蛋白受体 334
转座 56,463
转座酶 62
转座子 57,463
自交 4
自由核糖体 145
自主复制序列 48,463
自主性因子 59
阻遏蛋白 232,463
组氨酸 9
组成型剪接 287
组蛋白 22,463
组蛋白基因家族 284
组织相容性复合体 397,463
组织相容性抗原 384
左剪接位点 287
左手螺旋 33

[General Information]
书名=现代分子生物学
作者=朱玉贤，李毅，郑晓峰编著
页数=476
出版日期=2007.11
SS号=11899077